Population-based survival analysis of patients with cancer of the oral cavity in Finland in 1994-2014

Master's thesis in Statistics
Iiro Nerg
Research Unit of Mathematical Sciences
University of Oulu
Spring 2020
Supervisor: Professor Esa Läärä

Abstract

This population-based study describes the survival of patients with cancer of the oral cavity, diagnosed from 1994 to 2013 in Finland, followed up through 2014. The life table method and the relative survival framework are used to estimate net survival for the cancer patients under study. Two recommended estimators, the Ederer II and the Pohar Perme, for net survival are compared.

The material used in this thesis is from the Finnish Cancer Registry and consists of 4211 cancer cases. 201 cancer cases were diagnosed only in autopsy and were excluded from the analyses, leaving 4010 observations that were used for the survival analysis.

Survival curves are presented for observed and relative survival estimates. Analyses were stratified by sex, patients' age, stage of cancer and calendar year at the time of diagnosis.

Patient's age and stage of cancer turned out to be important predictors of patient survival from this cancer, which was expected from literature. The study period was divided into two ten-year calendar periods by the calendar years of diagnoses. There was not a clear difference between the survival of these periods.

The Ederer II and the Pohar Perme estimators of net survival performed similarly when a large quantity of data was available, e.g. when estimating the five-year relative survival ratio for cancer of the oral cavity. Estimates differed when less data were available, e.g. when the relative survival ratio for longer follow-up periods was estimated or when the analysis was stratified by age and patients in the older age groups were considered. In these situations the standard error of the Pohar Perme estimator was considerably larger than the standard error of the Ederer II estimator.

Tiivistelmä

Tässä väestöpohjaisessa tutkielmassa kuvaillaan suuontelon syöpäpotilaiden elinaikoja. Potilaat on diagnosoitu Suomessa vuosien 1994-2013 aikana ja heitä on seurattu vuoden 2014 loppuun. Tutkielmassa käytetään aktuaarimenetelmää ja suhteellista elossaololukua suuontelon syövän nettoelossaolon estimoimiseksi. Kahta suositeltua menetelmää, Ederer II ja Pohar Perme, verrattiin nettoelossaolon estimoimiseksi.

Tutkielmassa käytetty aineisto on Suomen Syöpärekisteriltä. Aineisto koostuu 4211:stä syöpäpotilaasta. Potilaista 201:llä oli todettu suuontelonsyöpä vasta ruumiinavauksessa, joten heidät poissuljettiin analyysista. Analyyseissa käytettävän aineiston koko oli lopulta 4010 potilasta.

Havaitun ja suhteellisen elossaolon estimaatit on esitelty välttökäyrien ja kiinnostavien tunnuslukujen avulla. Analyyseissä on käytetty osittavina tekijöinä sukupuolta, potilaan ikää, syövän levinneisyyttä ja kalenterivuotta diagnoosihetkellä. Vertailut tehdään kuvailevasti esitetyistä välttökäyristä ja taulukoiduista tunnusluvuista.

Potilaan ikä ja syövän levinneisyys ovat tärkeitä ennustetekijöitä, jotka vaikuttivat merkittävästi suuontelon syöpäpotilaan elinaikaan, mikä oli odotettavissa kirjallisuuden perusteella. Tutkimusjakso on jaettu kahteen kymmenvuotisjaksoon diagnoosivuosien perusteella. Jaksojen välillä ei ollut nähtävissä selvää eroa elinajoissa.

Nettoelossaolon Ederer II ja Pohar Perme -estimaatit olivat samansuuruisia, kun käytettävissä oli paljon aineistoa, esimerkiksi laskettaessa viiden vuoden suhteellista elossaoloa. Erot tulivat esiin kun käytettävissä oli vähemmän aineistoa, esimerkiksi laskettaessa suhteellista elossaoloa pitkille seuranta-ajoille tai iän mukaan ositetussa analyysissa vanhoilla ikäryhmillä. Kun aineistoa oli vähän, Pohar Perme -estimaatin keskivirhe kasvoi huomattavasti suuremmaksi kuin Ederer II -estimaatin keskivirhe.

Acknowledgements

This thesis started as a project work for the Nordic Summer School in Cancer Epidemiology 2017, held in Denmark and Finland. I am thankful to Finnish Cancer Registry for giving me this opportunity and providing the material for the study. This thesis has been a long time in the making and I am happy to finally complete this work.

First, I want to thank my supervisor at University of Oulu, Professor Esa Läärä, for his helpful comments on the work and his great patience working with me.

For the struggles and successes we shared, I am thankful to my friends, especially my old study group Päänhakkaajat, and my great buddies at ÄMK.

Finally, I am ever grateful for my family, who were always supportive of me.

"This world is a school—no one expects to stay in school forever"

Iiro Nerg
Oulu, 26.2.2020

Contents

1 Introduction

In population-based cancer survival studies, estimating net survival is a common practice. Net survival is used when the interest is the probability that a patient will die of a specific cause. Net survival cannot be estimated directly from the survival times of patients, thus a relative survival framework is commonly used. (Dickman and Coviello 2015; Rebolj Kodre and Pohar Perme 2013) Estimating net survival in a population-based setting with relative survival estimators is the key subject of this thesis.

Cancer of the oral cavity, together with cancer of the lip, is one of the more common of head and neck cancers in Finland. Common risk factors for this cancer are smoking and heavy alcohol use, separately causing a sixfold increase in the hazard of cancer and together increasing the hazard of cancer up to 15 times higher. Typically patients with cancer of the oral cavity are over 60 years old, but little over 10 % of patients are diagnosed before they are 40 years old. The cancer of the oral cavity tends to be quite agressive and can spread to metastases quite fast, sometimes a lump in the neck can even be the first symptom. Around 30 % of patients are diagnosed with metastases in the neck area, but metastases elsewhere in the body are rare. (Joensuu et al. 2013)

Patients diagnosed with a small tumor can have an expected five-year survival proportion up to 90 %, while for patients diagnosed with a large tumor it can be as low as 20 %. Furthermore, the five-year survival proportion is generally halved for patients who are diagnosed with a cancer that has spread to a metastasis on the same side of the neck, and reduced to a quarter for patients diagnosed with bilateral metastases, compared to those with local tumor. After a local recurrence or a discovery of metastasis, getting a curative treatment outcome is rare, and on average patients have a life expectancy less than one year. (Joensuu et al. 2013)

In Section 2 the methods of survival analysis are introduced. The focus is on the survival and the net survival functions, and their estimators. In Section 3 the material used in this thesis is described. In Section 4 the baseline characteristics of the cancer patients in the study population are presented first, then the results of survival analyses are presented. Finally The Ederer II and the Pohar Perme estimators of net survival are compared. Lastly in Section 5 the concept of net survival and results of the survival analyses are discussed.

2 Methods

In this chapter the characteristics of survival data and the methods of survival analysis are presented. Methods used in this thesis are focused. At the end of this chapter, the computational tools used to produce the results of this thesis are described.

2.1 Characteristics of survival data

Survival analysis is the analysis of data in the form of times from a well-defined *time origin* until an interesting *end-point*. In the medical field the end-point is often the occurence of a particular event, e.g. death, while the time origin corresponds to the recruitment of an individual into a study, commonly coinciding with the diagnosis of the particular condition. If the end-point is not fatal, observations can be referred to as *time to event data.*(Collett 2015)

Censoring is the main feature that renders standard data analysis methods inappropriate. Survival data are said to be censored when the interesting end-point is not observed. In this thesis, only *right censoring* is considered, which occurs after an individual has entered the study, to the right of the last known survival time. From now on it will be referred to as censoring. For more about different censoring processes see e.g. Leung, Elashoff, and Afifi (1997) and Collett (2015).

An individual is *lost to follow-up*, if they could not be reached for a check up, e.g. because emigration, and the only information available is the last day they are known to be alive. *Administrative censoring* happens when the observation period of a study ends or the data are analysed before the observation period ends, thus some of the patiens might be alive. (Leung, Elashoff, and Afifi 1997; Collett 2015)

Nonetheless, an individual enters the study at a time t_0 and dies at a time $t_0 + t$, but the time t is unknown. If the interesting end-point is not observed, but a patient was last known to be alive at a time $t_0 + c$, the time c is called a censored survival time. Most survival analysis methods require an assumption of *independent censoring*, to include censored survival times. This means that a time t must not be dependent on any mechanism that causes a survival time to be censored at a time c, where $c < t$. When the assumption is true, an individual is representative of all other individuals who have survived until the time c. (Collett 2015; Leung, Elashoff, and Afifi 1997)

The recruitment period of a study can extend over days, months or even years. This causes calendar time periods of individuals in the study, *study times*, to differ. The study begins for an individual when they are recruited at their time origin t_0, and in the medical field the period of time they spend in the study is often called *patient time*. (Collett 2015)

2.2 Survival, hazard and cumulative hazard functions

In this section the survival function, the hazard function and the cumulative hazard function are introduced. They are commonly used to describe the distribution of survival times. Some useful relationships between these functions are also derived. The theory in this sections is adopted from Collett (2015), Seppä (2012) and Seppä et al. (2016).

Let T be a random variable describing survival time and t be any fixed survival time, i.e. a possible realised value of T. Clearly t can only have non-negative values. Now, suppose T has some underlying *probability density function* $f(t)$, then *the cumulative distribution function* of T is

$$F(t) = P(T \leq t) = \int_0^t f(u)\mathrm{d}u. \tag{1}$$

This function is sometimes called *the cumulative incidence function*, because it describes the cumulative probability of the interesting end-point occuring before the time t.

Now, *the survival function* $S(t)$, is defined as the probability of the survival time being greater than t:

$$S(t) = P(T > t) = 1 - F(t). \tag{2}$$

An alternative presentation for the survival funtion is

$$S(t) = P(T > t) = \exp\left\{-\int_0^t h(u)\mathrm{d}u.\right\}, \tag{3}$$

where $h(t)$ is *the hazard function*:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t < T \leq t + \Delta t \mid T > t)}{\Delta t}. \tag{4}$$

The hazard function is interpreted as the instantaneous rate at which death occurs, having not occurred before. The considered probability is that the random variable T lies between times t and $t+\Delta t$, conditional on the survival time, T, being greater than t. The hazard function is then defined as the limiting value of this probability, divided by the time interval Δt, when the interval tends to zero.

If the hazard can be assumed constant over a short time period, say Δt, the expected number of events experienced in an unit time is $h(t) \times \Delta t$, given the event has not occurred before.

The cumulative hazard, $H(t)$, is given as

$$H(t) = \int_0^t h(u)\mathrm{d}u. \tag{5}$$

The cumulative hazard function summarises the hazard of an event of interest up to a time t, given the event has not occurred before. It is possible for the cumulative hazard function to exceed unity, which means that the expected number of events is greater than one in the time interval $(0,\ t)$. The interpretation that the cumulative hazard function is the expected number of events in a time interval is only reasonable if a repetition of an event is possible, e.g. an occurence of an infection or backpain.

Some useful relationships can be obtained between these functions. Note that the probability of an event A, given that an event B occurs, is $P(A|B) = P(A \cap B)/P(B)$, where $P(A \cap B)$ is the probability of the joint occurence of A and B. With this result, the probability in the Equation 4 can be written as

$$\frac{P(t < T \leq t + \Delta t)}{P(T > t)},$$

which is

$$\frac{F(t + \Delta t) - F(t)}{S(t)},$$

where $F(t)$ is the distribution function of T given in the Equation 1. Now, the hazard function can be written as

$$h(t) = \lim_{\Delta t \to 0} \left\{ \frac{F(t + \Delta t) - F(t)}{\Delta t} \right\} \frac{1}{S(t)},$$

where

$$\lim_{\Delta t \to 0} \left\{ \frac{F(t+\Delta t) - F(t)}{\Delta t} \right\} = F'(t) = f(t)$$

is the density function of T. Thus,

$$h(t) = \frac{f(t)}{S(t)}. \tag{6}$$

Also, the Equation 3 can be written as

$$S(t) = \exp\{-H(t)\}, \tag{7}$$

then we get

$$H(t) = -\log S(t). \tag{8}$$

and

$$h(t) = -\frac{\mathrm{d}}{\mathrm{d}t} \log S(t). \tag{9}$$

There are multiple ways to determine any of the presented functions. From any one of the four functions $f(t)$, $S(t)$, $h(t)$ and $H(t)$, the other three can be determined.

Considering these functions on an individual level, let T_i be a random variable representing the survival time of an individual i. The survival probability of an individual i is defined as

$$S_i(t) = P(T_i > t) = \exp\left\{ -\int_0^t h_i(u) \mathrm{d}u. \right\},$$

where $h_i(t)$ is the hazard of an individual i:

$$h_i(t) = \lim_{\Delta t \to 0} \frac{P(t < T_i \leq t + \Delta t \mid T_i > t)}{\Delta t}.$$

In a group on n individuals, the overall survival is the mean of the survival probabilities of the individuals in the group:

$$S(t) = \frac{1}{n} \sum_{i=1}^{n} S_i(t).$$

2.3 Net and relative survival

In this section net survival and relative survival, the latter as an estimate for net survival, are considered along with the concept of competing risks. The theory in this section is based on Seppä (2012) and Seppä et al. (2016).

Multiple causes of death are often present when survival data are analysed, though the interest usually is focused on a particular cause of death, e.g. cancer. Let C be a random variable related to the cause of death. *The cause-specific hazard*, $h_c(t)$, is defined as the instantaneous rate at which death occurs due to the cause $C = c$, given that death from any cause has not occurred before:

$$h_c(t) = \lim_{\Delta t \to 0} \frac{\mathrm{P}(t < T \leq t + \Delta t \ \cap \ C = c \mid T > t)}{\Delta t} \tag{10}$$

Given that different causes of death are mutually exclusive, the sum of the cause-specific hazards is equal to the overall hazard. For example, if $C = \{1, 2\}$, then $h(t) = h_1(t) + h_2(t)$.

When different competing risks of death are present, the cause-specific cumulative incidence function from the time of diagnosis to a time t is

$$F_c(t) = \mathrm{P}(T \leq t \ \cap \ C = c) = \int_0^t h_c(u) S(u) \mathrm{d}u.$$

Again, let T_i be a random variable representing the survival time of an individual i until death from any cause. Introducing the concept of competing risk to this variable, suppose T_i is the minimum of two random variables, $\min\{T_i^E, T_i^V\}$. Now, T_i^E is the time to death from a given cancer, and T_i^V is the time to death from other causes than the cancer. Only one of these survival times can be observed for each individual. If the competing risks are mutually independent, *the net survival function* could be defined as the survival probability of an individual i, in the absence of other causes of death, i.e. the hypothetical situation where other causes of death than cancer are eliminated:

$$S_{Ei}(t) = \mathrm{P}(T_i^E > t) = \exp\left\{-\int_0^t h_{Ei}(u)\mathrm{d}u\right\},$$

where $h_{Ei}(t)$ is the net hazard of the ith individual:

$$h_{Ei}(t) = \lim_{\Delta t \to 0} \frac{\mathrm{P}(t < T_i^E \leq t + \Delta t \mid T_i^E > t)}{\Delta t}.$$

In a group of n patients, the overall net survival is the mean of the net survival probabilities of the individuals in the group:

$$S_E(t) = \frac{1}{n}\sum_{i=1}^{n} S_{Ei}(t). \tag{11}$$

Additionally, if T_i^E and T_i^V can be assumed to be independent:

$$S(t) = S_E(t) \times S_V(t).$$

Thus the interesting net survival $S_E(t)$ could be estimated based on the overall survival $S(t)$ and the net survival $S_V(t)$. The latter refers to the hypothetical situation where deaths due to the interesting cancer are eliminated.

The probability interpretation of net survival requires the assumption about the independence of competing risks to be true. This assumption should be taken with a caution, since it can rarely be verified through data. From the Equation 11, the observable net survival can be defined by replacing the net hazard of an individual i, $h_{Ei}(t)$, with the observed excess hazard of said individual

$$h_{Ei}^*(t) = \lim_{\Delta t \to 0} \frac{\mathrm{P}(t < T_i^E \leq t + \Delta t \mid T_i > t)}{\Delta t}.$$

Relative survival offers a method for estimating net survival, even when the causes of death are unknown or assumed to be unreliable. The net survival if deaths due to cancer were eliminated, $S_V(t)$, is estimated by the expected survival function $S_P(t)$, that refers to the survival of a group, that is free of cancer, in a relevant reference population similar enough with the group under study, in terms of traits affecting mortality.

Expected survival can be estimated from a large national population. When stratified by sex, age and calendar year, the expected survival can be considered fixed, or basically free from random error. A reference population of a large nation obviously also contains cancer cases of the interest, and possibly includes individuals who are under study. This effect can be assumed negligible on the estimated relative survival of a large national population.

As stated before, if the assumption of the independence of T_i^E and T_i^V is invalid, the net survival has no meaningful probability interpretation. However, the relative survival ratio can always be interpreted as the ratio between the corresponding probabilities of an individual under study, and a healthy individual in the reference population, being alive at a time t, i.e. the ratio between observed and expected survival:

$$S_R(t) = \frac{S(t)}{S_P(t)}.$$

Furthermore, the excess hazard can be used as a surrogate estimate for the net hazard. It is the excess rate of death an individual with cancer has, compared to the rate of death in a comparable healthy person, $h_P(t)$:

$$h_R(t) = h(t) - h_P(t).$$

In a group of n individuals, the relative survival ratio is defined as the ratio between the averages of the observed and the expected patient-specific survival probabilities:

$$S_R(t) = \frac{\sum_{i=1}^n S_i(t)}{\sum_{i=1}^n S_{Pi}(t)}. \tag{12}$$

However, in a group of n individuals the net survival is the average of the patient-specific net survival probabilities $S_{Ei}(t)$ as defined in the Equation 11. Now, the net survival of an individual i can be written as the ratio between the observed and the expected survival of said individual:

$$S_E(t) = \frac{1}{n}\sum_{i=1}^n S_{Ei}(t) = \frac{1}{n}\sum_{i=1}^n \frac{S_i(t)}{S_{Pi}(t)}. \tag{13}$$

2.4 Life table estimators

In this section the life table method, and estimators for the survival function, the hazard function and the cumulative hazard function based on it are presented. The theory in this section is based on Collett (2015), Gehan (1969), Seppä (2012) and Seppä et al. (2016).

Two popular non-parametric approaches to estimating the survival function are *the Kaplan-Meier method* and *the life table method*. The latter is preferred in this thesis, since exact survival times must be observed to use the Kaplan-Meier estimates, and the population-based survival data used in this thesis does not have the exact dates. Also in such cases when the exact survival times are known, life table estimates can still be used, but might lead to some loss of information. Alternative methods, such as the Kaplan-Meier, are then more appropriate, see e.g. Collett (2015).

In the life table method, sometimes also called *the actuarial method*, the survival data are grouped by dividing the period of observation into a series of disjoint time intervals. These intervals need not to be of same length, and the number of intervals can depend on the number of individuals in the study.

Consider the jth of such J intervals: $[t_{j-1}, t_j)$, $j = 1, 2, \ldots, J$, where $t_0 = 0$ and $t_J = \infty$. Let n_j be the number of individuals alive and under follow-up at the start of the jth interval, thus being at risk of death. Let d_j be the number of deaths and c_j be the number of individuals with a censored survival time in the interval j. *The actuarial assumption*, is that censoring happens uniformly throughout the jth interval and independently of death, i.e. individuals with a censored survival time are assumed to be at risk of dying for half of the duration of the interval, on the average. The average number of individuals at risk during the jth interval, i.e. the effective denominator, is

$$n'_j = n_j - \frac{c_j}{2}. \tag{14}$$

The estimated probability of death in the jth interval is $\hat{q}_j = d_j/n'_j$, and the estimated survival probability over the interval is

$$\hat{p}_j = 1 - \hat{q}_j = \frac{n'_j - d_j}{n'_j}. \tag{15}$$

If there are no censored survival times, the survival function $S(t)$, defined in the Equation 2, is simply estimated by *the empirical survival function*:

$$\tilde{S}(t) = 1 - \tilde{F}(t) = \frac{\text{Number of individuals with a survival time} \geq t}{\text{Number of individuals in the data}}, \quad (16)$$

where $\tilde{F}(t)$ is *the empirical distribution function*, i.e. the ratio of individuals alive at a time t to the total number of individuals. $\tilde{S}(t)$ is unity for the values of t before the first observed death time and zero after the last observed death time, and it is assumed to be constant between adjacent death times.

When there are censored survival times, the probability that an individual survives beyond a time t_{j-1}, $j = 1, 2, \ldots, J$, i.e. beyond the start of the jth interval, is the product of the conditional probabilities that an individual survives through all the preceding $j - 1$ intervals. The life table estimate of the survival function is then

$$\hat{S}(t) = \prod_{j=1}^{k} \left(\frac{n'_j - d_j}{n'_j} \right) = \prod_{j=1}^{k} (1 - \hat{q}_j) = \prod_{j=1}^{k} \hat{p}_j, \quad (17)$$

for $t_{j-1} \leq t < t_j$, $j = 1, 2, \ldots, J$. The probability of surviving beyond the start of the first interval, t_0, is unity, while the probability of surviving beyond t_J is zero.

The hazard function is defined in the Equation 4. To estimate it, a constant death rate is assumed during the jth interval. An appropriate estimate of the average hazard of death per unit time, is the observed deaths divided by the average time survived, in the jth interval. Assuming the constant death rate, the average time survived during an interval is $(n'_j - d_j/2)\tau_j$, where τ_j is the length of the jth interval. The estimated hazard function, $\hat{h}(t)$, is a step-function over the observation period, and in the jth interval estimated by

$$\hat{h}(t) = \frac{2\hat{q}_j}{(1 - \hat{p}_j)\tau_j} = \frac{d_j}{(n'_j - d_j/2)\tau_j}, \quad (18)$$

for $t_{j-1} \leq t < t_j$, $j = 1, 2, \ldots, J$.

The cumulative hazard at a time t, $H(t)$, is defined in the Equation 5 to be the hazard function integrated until t. However, it is more convenient to find the cumulative hazard function using the Equation 8. $\hat{S}(t)$, given in the Equation 17, is the life table estimate of the survival function and an appropriate estimate of the cumulative hazard function is

$$\hat{H}(t) = -\log \hat{S}(t),$$

for $t_{j-1} \leq t < t_j,\ j = 1, 2, \ldots, J$. Since the derivative of the cumulative hazard function is the hazard function, the slope of the cumulative hazard function provides information about the underlying hazard function. e.g. a linear cumulative hazard function over some time interval suggests that the hazard would be constant over that interval (Collett 2015).

2.5 Standard error and confidence interval

This section considers the standard error and the confidence interval of the life table estimator of the survival function. For the derivation of the standard error of the life table estimate of the hazard function, see Gehan (1969). The theory in this section is based on Collett (2015), Gehan (1969), Seppä (2012) and Seppä et al. (2016).

To derive the standard error of the survival function, we start by taking logarithms in the Equation 17

$$\log \hat{S}(t) = \sum_{j=1}^{k} \log \hat{p}_j,$$

and the variance of $\log \hat{S}(t)$ is

$$\text{var}\left\{\log \hat{S}(t)\right\} = \sum_{j=1}^{k} \text{var}\{\log \hat{p}_j\}. \tag{19}$$

Now, a binomial distribution is assumed for the individuals who survive through the interval j. The binomial distribution takes parameters n'_j and p_j, where the latter is the true probability to survive through the jth interval. Using the result that the variance of a binomial random variable with parameters n and p is $np(1-p)$, the variance of the observed number of individuals who survive through the jth interval, $n'_j - d_j$, is

$$\text{var}(n'_j - d_j) = n'_j p_j (1 - p_j).$$

From the Equation 15

$$\text{var}\left(\hat{p}_j\right) = \frac{\text{var}(n'_j - d_j)}{(n'_j)^2} = \frac{p_j(1-p_j)}{n'_j}.$$

Thus, the variance of $\hat{p}_j$ is estimated by

$$\widehat{\text{var}}\left(\hat{p}_j\right) = \frac{\hat{p}_j(1-\hat{p}_j)}{n'_j}. \tag{20}$$

To obtain the variance of $\log \hat{p}_j$, we use the general result that the approximate variance of a funtion $g(X)$ of the random variable X is

$$\text{var}\{g(X)\} \approx \left\{\frac{\text{d}g(X)}{\text{d}X}\right\}^2 \text{var}(X). \tag{21}$$

This is known as *the Taylor series approximation* of the variance of a function of a random variable. With the Equation 21:

$$\text{var}\left(\log \hat{p}_j\right) \approx \frac{\text{var}(\hat{p}_j)}{\hat{p}_j^2}.$$

Then, with the Equation 20, and a substitution for $\hat{p}_j$:

$$\widehat{\text{var}}\left(\log \hat{p}_j\right) \approx \frac{1-\hat{p}_j}{n'_j \hat{p}_j} = \frac{d_j}{n'_j(n'_j - d_j)}. \tag{22}$$

Applying this to the Equation 19

$$\widehat{\text{var}}\left\{\log \hat{S}(t)\right\} \approx \sum_{j=1}^{k} \frac{d_j}{n'_j(n'_j - d_j)}, \tag{23}$$

and further applying the result in the Equation 21:

$$\widehat{\text{var}}\left\{\log \hat{S}(t)\right\} \approx \frac{1}{\left[\hat{S}(t)\right]^2} \text{var}\left\{\hat{S}(t)\right\}.$$

So the approximated variance of the survival function is

$$\widehat{\text{var}}\left\{\hat{S}(t)\right\} \approx \left[\hat{S}(t)\right]^2 \sum_{j=1}^{k} \frac{d_j}{n'_j(n'_j - d_j)}. \tag{24}$$

The standard error is defined as the square root of the variance of the estimate, so the approximated standard error of the life table estimate of the survival function is

$$\operatorname{se}\left\{\hat{S}(t)\right\} \approx \hat{S}(t)\left\{\sum_{j=1}^{k}\frac{d_j}{n'_j(n'_j-d_j)}\right\}^{\frac{1}{2}}, \tag{25}$$

for $t_{j-1} \leq t < t_j$. This result is known as *the Greenwood's formula*.

If there is no censored survival times in the sample, i.e. $n_{j+1} = n_j - d_j$, the Equation 22 can be reduced:

$$\widehat{\operatorname{var}}\left(\log \tilde{p}_j\right) \approx \frac{n_j - n_{j+1}}{n_j n_{j+1}},$$

and

$$\operatorname{var}\left\{\log \tilde{S}(t)\right\} \approx \sum_{j=1}^{k}\frac{n_j - n_{j+1}}{n_j n_{j+1}} = \sum_{j+1}^{k}\left(\frac{1}{n_{j+1}} - \frac{1}{n_j}\right) = \frac{n_1 - n_{k+1}}{n_1 n_{k+1}}. \tag{26}$$

Since $\tilde{S}(t) = n_{k+1}/n_1$ for $t_{j-1} \leq t < t_j,\ j = 1, 2, \ldots, J$, in the absence of censoring, the Equation 26 can be written as

$$\frac{1-\tilde{S}(t)}{n_1\tilde{S}(t)}.$$

Thus, from the Equation 24

$$\widehat{\operatorname{var}}\left\{\tilde{S}(t)\right\} \approx \frac{\tilde{S}(t)\left[1-\tilde{S}(t)\right]}{n_1}.$$

This is an estimate for the variance of the empirical survival function in the Equation 16, assuming the number of individuals at risk at a time t has a binomial distribution with parameters n_1, $S(t)$.

A confidence interval (CI) is an interval estimate, with a prescribed probability that the true value of the estimate is covered by the random interval. The presented intervals are *pointwise confidence intervals*, since they refer to a specific time point, and can be found at any given time t. Assuming the estimated value of the survival function, at a given time t, to be normally distributed with a mean $S(t)$, and an estimated variance given by the Equation

24, a confidence interval for the true value of the survival function $S(t)$ at a time t can be computed from *percentage points* of the standard normal distribution.

Now, if Z is a random variable with a stardard normal distribution, the upper (one-sided) $\alpha/2$-point of the distribution is the value $z_{\alpha/2}$, such that $\mathrm{P}(Z > z_{\alpha/2}) = \alpha/2$. Then a $100(1-\alpha)\%$ approximate confidence interval for $S(t)$, for a given value of t, is the interval

$$\left[\hat{S}(t) - z_{\alpha/2} \times \mathrm{se}\{\hat{S}(t)\},\ \hat{S}(t) + z_{\alpha/2} \times \mathrm{se}\{\hat{S}(t)\}\right],$$

where $\mathrm{se}\{\hat{S}(t)\}$ is from the Equation 25.

Few problems may arise with this procedure, one being the symmetry of the confidence interval. When $\hat{S}(t)$ is close to zero or unity, the symmetric confidence limits can lie outside the interval $(0, 1)$, making them inappropriate. A simple solution is to replace any limit exceeding unity by 1.0 and any limit below zero by 0.0.

Alternatively, $\hat{S}(t)$ can be transformed to a value in the range $(-\infty,\ \infty)$. A confidence interval is obtained for the transformed value and back-transformed to give a confidence interval for $S(t)$. Possible transformations are the logistic transformation, $\log[S(t)/\{1 - S(t)\}]$, and the complementary log-log transformation, $\log\{-\log S(t)\}$. Now, the variance of the latter quantity is obtained from the Equation 23 by using the general result in the Equation 21:

$$\mathrm{var}\{\log(-X)\} \approx \frac{1}{X^2}\mathrm{var}(X).$$

Setting $X = \log \hat{S}(t)$ gives

$$\mathrm{var}\left\{\log[-\log \hat{S}(t)]\right\} \approx \frac{1}{\{\log \hat{S(t)}\}} \sum_{j=1}^{k} \frac{d_j}{n'_j(n'_j - d_j)}. \tag{27}$$

Now the square root of the Equation 27 is the standard error of $\log\{-\log S(t)\}$, which leads to a $100(1 - \alpha)\%$ approximate confidence interval limits for $S(t)$ of the form

$$\hat{S}(t)^{\exp[\pm z_{\alpha/2} \times \mathrm{se}\{\log[-\log \hat{S}(t)]\}]}$$

Another problem is that the variance of $\hat{S}(t)$, obtained using the Greenwood's formula, can underestimate the actual variance, when $\hat{S}(t)$ is close to unity

or zero in the tails of the distribution of the survival times. In these circumstances, an alternative expression for the standard error of $\hat{S}(t)$ may be used, but the Greenwood estimate is recommended for general use (Collett 2015).

2.6 Estimating the median and percentiles of survival times

In this section a method for estimating the median, and other percentiles, of survival times is presented. The theory in this section is based on Collett (2015), Gehan (1969), Seppä (2012) and Seppä et al. (2016).

The median survival time is often the preferred summary measure of the location of the distribution of survival times. The distribution of survival times tends to be positively skewed and obtaining the mean survival time, more often than not, requires an extrapolation of the survival times. The median survival time is obtained from the estimated survival function, it is the time beyond which 50 % of the individuals in the population under study are expected to survive. This is the value $t(50)$ such that $S\{t(50)\} = 0.5$. Since the life table estimate of the survival function is a non-parametric step-function, it is rare to realise the exact value. Instead, the estimated median survival time, $\hat{t}(50)$, is the shortest observed survival time for which the estimated survival function is less than 0.5:

$$\hat{t}(50) = \min\{t \mid \hat{S}(t) < 0.5\}.$$

In a situation where the estimated survival function is exactly equal to 0.5, for $t_{j-1} \leq t < t_j$, the median is taken at the halfway point of that interval, $(t_{j-1}+t_j)/2$. Without any censored survival times in the sample, the estimated median is the shortest survival time beyond which 50 % of the individuals survive.

A similar procedure can be used to estimate other *percentiles* of the distribution of survival times: the pth percentile is the value $t(p)$, such that $S\{t(p)\} = 1 - (p/100)$ for $p \in [0, 100]$. E.g. the 10th and the 90th percentiles are

$$S\{t(10)\} = 0.90 \quad \text{and} \quad S\{t(90)\} = 0.10.$$

With the estimated survival function, the estimated pth percentile is the smallest observed survival time, $\hat{t}(p)$, such that $\hat{S}\{\hat{t}(p)\} < 1 - (p/100)$. In some

cases, the estimated survival function is greater than 0.5 for all values of t and the median survival time can't be estimated. Then it is natural to summarise the distribution of survival times through other percentiles.

The variance of the estimated median and percentiles can be approximated by using the result in the Equation 21:

$$\text{var}[\hat{S}\{\hat{t}(p)\}] \approx \left(\frac{\text{d}\hat{S}\{\hat{t}(p)\}}{\text{d}\hat{t}(p)}\right)^2 \text{var}\{\hat{t}(p)\}, \tag{28}$$

where

$$-\frac{\text{d}\hat{S}\{\hat{t}(p)\}}{\text{d}\hat{t}(p)} = \hat{f}\{\hat{t}(p)\}$$

is an estimate of the probability density function of survival times at $\hat{t}(p)$. Now, rearranging the Equation 28 gives the approximated variance of the estimated percentile p:

$$\text{var}\{\hat{t}(p)\} \approx \left(\frac{1}{\hat{f}\{\hat{t}(p)\}}\right)^2 \text{var}[\hat{S}\{\hat{t}(p)\}].$$

Thus, the standard error of the estimated pth percentile, is given by

$$\text{se}\{\hat{t}(p)\} \approx \frac{1}{\hat{f}\{\hat{t}(p)\}}\text{se}[\hat{S}\{\hat{t}(p)\}]. \tag{29}$$

The Greenwood's formula given in the Equation 25 is used to find $\text{se}[\hat{S}\{\hat{t}(p)\}]$ and

$$\hat{f}\{\hat{t}(p)\} = \frac{\hat{S}\{\hat{u}(p)\} - \hat{S}\{\hat{l}(p)\}}{\hat{l}(p) - \hat{u}(p)},$$

where

$$\hat{u}(p) = \max\left\{t|\hat{S}(t) \geq 1 - \frac{p}{100} + \epsilon\right\},$$

and

$$\hat{l}(p) = \min\left\{t|\hat{S}(t) \leq 1 - \frac{p}{100} - \epsilon\right\},$$

for small values of ϵ. Usually, $\epsilon = 0.05$ is satisfactory, but greater values are needed if $\hat{u}(p)$ and $\hat{l}(p)$ turn out to be equal. Once $\text{se}\{\hat{t}(p)\}$ has been found, a $100(1-\alpha)\%$ confidence interval for $\hat{t}(p)$ has limits of

$$\hat{t}(p) \pm z_{\alpha/2} \times \text{se}\{\hat{t}(p)\}.$$

This interval estimate is only approximate, in the sense that the probability that the true percentile lies within the interval may not be exactly $1-\alpha$. Other methods have been proposed for constructing the confidence interval for the median and percentiles, but will not be presented in this thesis (Collett 2015).

2.7 Estimating net survival

In this section, two methods for estimating net survival will be presented, the Pohar Perme (PP) and the Ederer II (E2) method. Their variances and a method for age-standardisation are also given. In this thesis, other well-known methods are left out (e.g. Ederer I and Hakulinen), since the two presented methods have been recently recommended (Seppä et al. 2016). Only life table estimators are given. The theory in this section is based on Seppä et al. (2016). Ederer I and Hakulinen methods of estimating net survival are described in Dickman and Coviello (2015).

The Pohar Perme estimator is based on the weighted individual-level observations of n_j, d_j and c_j:

$$n_j^w = \sum_{i=1}^{n} \frac{n_{ij}}{S_{Pij}}, \quad d_j^w = \sum_{i=1}^{n} \frac{d_{ij}}{S_{Pij}} \text{ and } c_j^w = \sum_{i=1}^{n} \frac{c_{ij}}{S_{Pij}}, \tag{30}$$

where S_{Pij} is the patient-specific cumulative expected survival probability, calculated at the mid point of the interval j, $\bar{t}_j = (t_{j-1} - t_j)/2$. It is calculated with the interval-specific expected survival probabilities p_{ij}^* $(j = 1, \ldots, J)$ from national population life tables:

$$S_{Pik} = \prod_{j=1}^{J-1} p_{ij}^* \sqrt{p_{iJ}^*},$$

where $\sqrt{p_{iJ}^*}$, the conditional survival probability, is assumed to be equal for the first and the second half of the Jth interval.

The Pohar Perme etsimator of the net survival probability is

$$\hat{S}_E^{PP}(t) = \prod_{j=1}^{k} \left(1 - \frac{d_j^w}{n_j^w - c_j^w/2}\right) \Big/ \exp\left(-\frac{\tilde{d}_j^{*w}}{n_j^w - c_j^w/2 - d_j^w/2}\right), \qquad (31)$$

for $t_{j-1} \leq t < t_j, j = 1, 2, \ldots, J$, where $\tilde{d}_j^{*w}$ is the expected number of deaths in the interval j weighted by the cumulative expected survival probabilities:

$$\tilde{d}_j^{*w} = \sum_{i=1}^{n} \frac{-\log\left(p_{ij}^*\right)(n_{ij} - c_{ij}/2 - d_{ij}/2)}{S_{Pij}}.$$

Note that for an individual i in the interval j the follow-up time is $\Delta_j = t_j - t_{j-1}$, if the individual is alive at the end of the interval. If the individual dies or is censored during the interval, the follow-up time is half of the length of the interval $\Delta_j/2$.

The estimated variance of the Pohar Perme estimator is

$$\widehat{\text{var}}\left[\hat{S}_E^{PP}(t)\right] = \left[\hat{S}_E^{PP}(t)\right]^2 \sum_{j=1}^{k} \frac{\sum_{i=1}^{n} d_{ij}/S_{Pij}^2}{\left(n_j^w - c_j^w/2 - d_j^w/2\right)^2}.$$

Here, the weighted number of person-years in the interval j is approximated with $\left(n_j^w - c_j^w/2 - d_j^w/2\right)\Delta_j$. Thus the estimator is modified from the estimated variance of the Pohar Perme estimator of net survival derived with the hazard approach (Seppä et al. 2016).

The Ederer II estimator is a special case of the Pohar Perme estimator in the Equation 31. Here the same interval-specific weight is used for each individual, $S_{Pij} = S_{Pj}$, for all i:

$$\hat{S}_E^{E2}(t) = \prod_{j=1}^{k} \left(1 - \frac{d_j}{n_j - c_j/2}\right) \Big/ \exp\left(-\frac{\tilde{d}_j^*}{n_j - c_j/2 - d_j/2}\right),$$

where

$$\tilde{d}_j^* = \sum_{i=1}^{n} -\log\left(p_{ij}^*\right)\ (n_{ij} - c_{ij}/2 - d_{ij}/2)$$

is the expected number of deaths in the interval j. The estimated variance is

$$\widehat{\text{var}}\left[\hat{S}_E^{E2}(t)\right] = \left[\hat{S}_E^{E2}(t)\right]^2 \sum_{j=1}^{k} \frac{d_j}{(n_j - c_j/2)\ (n_j - c_j/2 - d_j)}.$$

This estimator is based on the Greenwood's formula on the variance of the cumulative observed survival proportion (Seppä et al. 2016). In a case that all individuals that enter an interval, die during the interval, both of these point estimators and their variances become zero.

Internal age standardisation is a common practise when estimating net survival, because net survival, and the background mortality of the general population usually depend on the same demographic factors. Old patients have a higher risk of dying from causes other than cancer, thus their contribution to long-term net survival is underweighted without age standardisation. The age distribution of patients at the time of diagnosis is used as the standard. In the life table method age standardisation is relevant for both of the presented estimates, for it also controls for the effect of informative censoring, that is due to the heterogenity in the potential follow-up times. Additionally, for the Pohar Perme estimator, age standardisation adjusts for the differences in the age distribution between individuals diagnosed prior to, and within, the period window.

Individuals are divided into, say M, age groups according to their age at time of diagnosis. Now, let $n_1(m)$ be the number of individuals alive and under observation at the beginning of the follow-up in age group m. Both of the presented estimators, PP an $E2$, are weighted averages of the pertinent age-specific estimators $\hat{S}_E(t,m),\ m = 1, 2, \ldots, M$:

$$\hat{S}_E(t) = \frac{1}{n_1} \sum_{m=1}^{M} n_1(m) \hat{S}_E(t,m).$$

The estimated variance of the age-standardised estimate of net survival:

$$\widehat{\text{var}}\left[\hat{S}_E(t)\right] = \frac{1}{n_1^2} \sum_{m=1}^{M} [n_1(m)]^2 \widehat{\text{var}}\left[\hat{S}_E(t,m)\right].$$

The estimate is unavailable in the last follow-up time interval, if the follow-up is incomplete and some individuals in an age group survive over the interval.

The confidence interval of net survival can be approximated, for each estimator with its variance, on the logarithmic scale:

$$\exp\left\{\log\left(\hat{S}_E(t)\right) \pm z_{\alpha/2} \times \widehat{\text{var}}\left[\log\left(\hat{S}_E(t)\right)\right]\right\},$$

where the variance of the natural logarithm of the net survival probability is approximated with the delta method:

$$\widehat{\text{var}}\left[\log\left(\hat{S}_E(t)\right)\right] = \left[\hat{S}_E(t)\right]^{-2} \widehat{\text{var}}\left[\hat{S}_E(t)\right],$$

and $z_{\alpha/2}$ is the $100(1-\alpha/2)$ percentile point of the standard normal distribution.

2.8 Computational tools

To produce the results of this thesis, the statistical computing language R version 3.6.1 (R Core Team 2019) was used in an integrated development environment RStudio version 1.1.456 (RStudio Team 2016).

The *Lexis* function from the *Epi* package (Carstensen et al. 2019) was used to transform the follow-up times into a Lexis type object, which is needed to estimate the survival time functions. The *survtab* function from the *popEpi* package (Miettinen and Rantanen 2019) was used to estimate the survival functions presented.

The *survtab* function estimates the observed survival as well as the Ederer II and the Pohar Perme estimators of net survival. The life table methods of estimation, described in Section 2, were used. Breaks in the life table intervals were set at one year apart after diagnosis. The complementary log-log transformation and the delta method are used to calculate the confidence interval for the survival function by default (Miettinen and Rantanen 2019).

To estimate the expected survival proportions, knowledge about the expected hazard levels in the population is required. This information was provided in the *popmort* dataset, included in the *popEpi* package, which contains the population mortality rates in Finland in 1951 - 2013 in 101 one-year age groups by gender (Miettinen and Rantanen 2019).

The package *colorspace* (Zeileis et al. 2019) was used for clear and colour blind friendly visualisation of the results, as suggested by Zeileis, Hornik, and Murrell (2009).

3 Material

The data used in this thesis were provided by Finnish Cancer Registry containing all cases of cancer of the oral cavity diagnosed in Finland between 1994 and 2013, a total of 4211 cases. Out of all cases, 201 were diagnosed only in autopsy and were excluded from the analyses, thus 4010 cases were used in the analyses. The patients in the data were born between 1902 and 1998 and the follow-up period extended through 2014.

The data consist of cancers with their primary location in the oral cavity, which consists of the tongue, tonsils, gums, the floor and the palate of mouth, the uvula, the retromolar area, the vestibule of mouth and an otherwise unspecified mouth area. Cancers of the lip and the pharynx area were not included. In Table 1 are listed all specific locations, ICD-O-3 codes (World Health Organization 2013) and frequencies.

The data are virtually free of missing observations of cancer cases, since all health organizations in Finland have a statutory obligation to provide information about every detected case or strong suspicion of cancer to the Finnish Cancer Registry. The coverage of all cancer cases in Finland is reported to be about 96 % (Finnish Cancer Registry 2020). The Finnish Cancer Registry receives data about cancer in Finland from institutions, e.g. hospitals, that provide treatment for cancer patients, healthcare professionals, pathology laboratories, and Statistics Finland provides data about causes of death. Information is received from multiple sources, and data from clinical notifications are considered to be the most accurate, if conflicts appear. The database of the registry is continuously updated and corrected, if errors are detected. (Finnish Cancer Registry 2020)

There are a few censored observations in the data, those are patients that have emigrated during the follow-up and patients that are still alive at the end of 2014, when the follow-up ended.

Few changes were done to the original data. The stage of cancer is originally provided as a variable with seven levels. In the analyses of this thesis “Local” and “Unknown” levels were kept as they are, but the remaining levels were combined into a “Spread” level. The status when patients exit the study was originally provided as a variable with four levels, but was combined in the analyses into two levels: dead with observed survival time or censored without observed survival time.

Birthdays of the patients were provided exactly, but for the date of diagnosis and the date of death or emigration only the month and the year are known.

Since the data were lacking these exact dates, the actuarial assumption was used: the date of diagnosis, and the date of death or emigration, was set to the middle point of the given month. For patients that were alive at the end of the data collection, the exit time was set to the closing date of the follow-up period.

Some patients in the data only had a record of the year of diagnosis. In these cases the actuarial assumption was extended and the date of diagnosis was set to the middle point of the year. The patients who did not have an exact record of the month of diagnosis, and death had occurred before the middle point of the same year, the month of diagnosis was set so, that the patient was under follow-up for half of the time they were alive that year. Patiens who were diagnosed and had died in the same month were set to have had a survival time of approximately two weeks.

Code	Description	N
C02.0	Dorsal surface of tongue, NOS	13
C02.1	Border of tongue	466
C02.2	Ventral surface of tongue, NOS	35
C02.3	Anterior 2/3 of tongue, NOS	16
C02.4	Lingual tonsil	12
C02.8	Overlapping lesion of tongue	6
C02.9	Tongue, NOS	1243
C03.0	Upper gum	63
C03.1	Lower gum	153
C03.9	Gum, NOS	31
C04.0	Anterior floor of mouth	50
C04.1	Lateral floor of mouth	20
C04.8	Overlapping lesion of floor of mouth	5
C04.9	Floor of mouth, NOS	820
C05.0	Hard palate	30
C05.1	Soft palate, NOS	52
C05.2	Uvula	12
C05.8	Overlapping lesion of palate	3
C05.9	Palate, NOS	61
C06.0	Cheek mucosa	152
C06.1	Vestibule of mouth	4
C06.2	Retromolar area	31
C06.8	Overlapping lesion of other and unspecified parts of mouth	5
C06.9	Mouth, NOS	726
Total		4010

Table 1: All topography codes in the data and their frequencies
Abbrevations: NOS – Not Otherwise Specified

4 Results

In this chapter the baseline characteristics and the results of the survival analyses are presented. The survival analyses are stratified by patients' sex, age, stage of cancer and calendar time at the time of diagnosis. The Ederer II estimator of net survival was used in the analyses. At the end of this chapter, the Pohar Perme and the Ederer II estimators of net survival are compared.

4.1 Baseline characteristics

In Table 2, the distribution of the cases of cancer used in the analyses of this thesis are presented by key variables. There were more men in the data, 52 %. Throughout the analyses sex is an important variable that is adjusted for.

	Men	(%)	Women	(%)	Total	(%)
Age at diagnosis						
0-44 years	170	(8)	137	(7)	307	(8)
45-54 years	430	(21)	213	(11)	643	(16)
55-64 years	649	(31)	373	(19)	1022	(25)
65-74 years	505	(24)	458	(24)	963	(24)
75-100 years	340	(16)	735	(38)	1075	(27)
Cancer stage						
Local	813	(39)	867	(45)	1680	(42)
Spread	842	(40)	557	(29)	1399	(35)
Unknown	439	(21)	492	(26)	931	(23)
Calendar year of diagnosis						
1994-1998	359	(17)	377	(20)	736	(18)
1999-2003	468	(22)	395	(21)	863	(22)
2004-2008	544	(26)	529	(28)	1073	(27)
2009-2013	723	(35)	615	(32)	1338	(33)
Total (%)	2094	(52)	1916	(48)	4010	(100)

Table 2: Diagnosed cases of cancer of the oral cavity in Finland in 1994-2013 by key variables.

The first variable in Table 2 is the patient's age at the time of diagnosis. These age groups are generally recommended, i.e. by Seppä et al. (2016). There was a considerable difference in the distribution of ages between sexes. More women

were diagnosed at the older ages. Most women were diagnosed when they were 75-100 years old. The largest age group for men was 55-64 years. More men than women were diagnosed at younger ages, 45-54 years and 55-64 years. In the whole population these effects were not present, all age groups after the age of 55 years had an approximately similar number of diagnosed cases of cancers of the oral cavity.

The second variable in Table 2 is the stage of cancer at the time of diagnosis. In the population, most cases were diagnosed as "Local", 42 %, while 35 % of the cases were diagnosed as "Spread". Between men and women, there was a difference of 11 percentage points in the cases of cancer that were diagnosed as "Spread". The majority of cancers for men, 40 %, were diagnosed as "Spread".

The third variable in Table 2 is the calendar year at the time of diagnosis. The period 1994-2013, when the patients were diagnosed, was divided into four five-year periods. The number of diagnosed cases of cancer is larger in the later five-year periods. There was little difference between men and women within these periods.

4.2 Overall survival

In Figure 1, the observed survival curve and the curve for the Ederer II estimator of net survival are presented, separately for men and women. In Table 3, values of both estimators are presented at five and ten years after diagnosis.

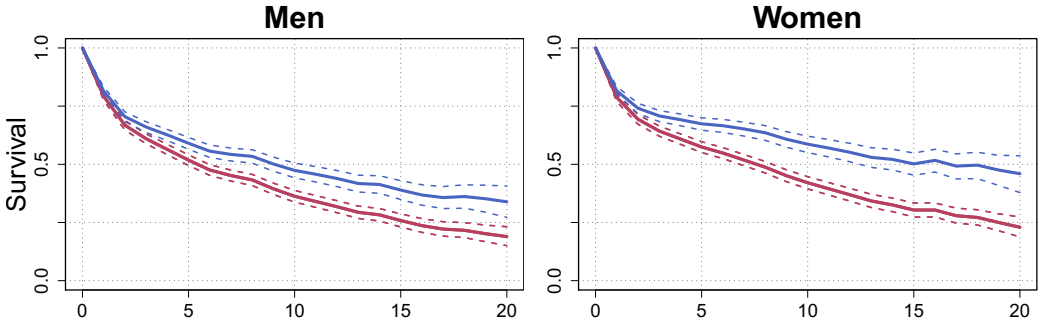

Figure 1: Comparison between the observed survival proportion and the Ederer II estimator of net survival. The red line corresponds to the observed survival proportion and the blue line corresponds to the Ederer II estimator. The dashed lines show the 95 % confidence intervals.

Time	Method	Sex			
		Men	(SE)	Women	(SE)
5 years	Observed	52	(1.2)	57	(1.2)
	Relative	59	(1.3)	67	(1.4)
10 years	Observed	36	(1.2)	42	(1.3)
	Relative	47	(1.6)	59	(1.8)

Table 3: The observed survival proportions per 100 and the Ederer II estimators of net survival per 100 at five and ten years after diagnosis.

The estimated net survival proportion was higher than the observed survival proportion at five and ten years after diagnosis when stratified by sex. At the same time, women had a higher observed and relative five-year and ten-year survival than men. When the analysis was stratified by sex the standard error of the estimates was reasonably low for both estimates for the whole follow-up period of 20 years.

As seen in Table 2, cancer affected older people more frequently, while older people in the population simultaneously have a higher overall hazard of death from all sources. This higher hazard weighs the hazard of death from cancer down, thus the estimated net survival was higher than the observed survival. As mentioned in Section 4.1, there were more women diagnosed at older age than there were men. The difference between the estimated net survival and the observed survival was bigger for women than for men.

In Table 4, the median survival times for men and women are presented. The observed median survival time had a stable estimate for both men and women in this sample. Women had approximately two years longer median survival time than men. The median survival time estimated with the Ederer II estimator was less precise, especially that for women had a large standard error.

Method	Sex			
	Men	(SE)	Women	(SE)
Observed	6	(0.3)	8	(0.4)
Relative	10	(0.8)	17	(2.7)

Table 4: Median survival times (years) calculated with the observed and the relative survival estimates.

4.3 Survival by age at the time of diagnosis

In Figure 2, the observed and the relative survival curves are presented, stratified by patient's age at the time of diagnosis. In Table 5, the estimates of the observed and the relative survival function at 5 years after diagnosis are presented in the different age groups.

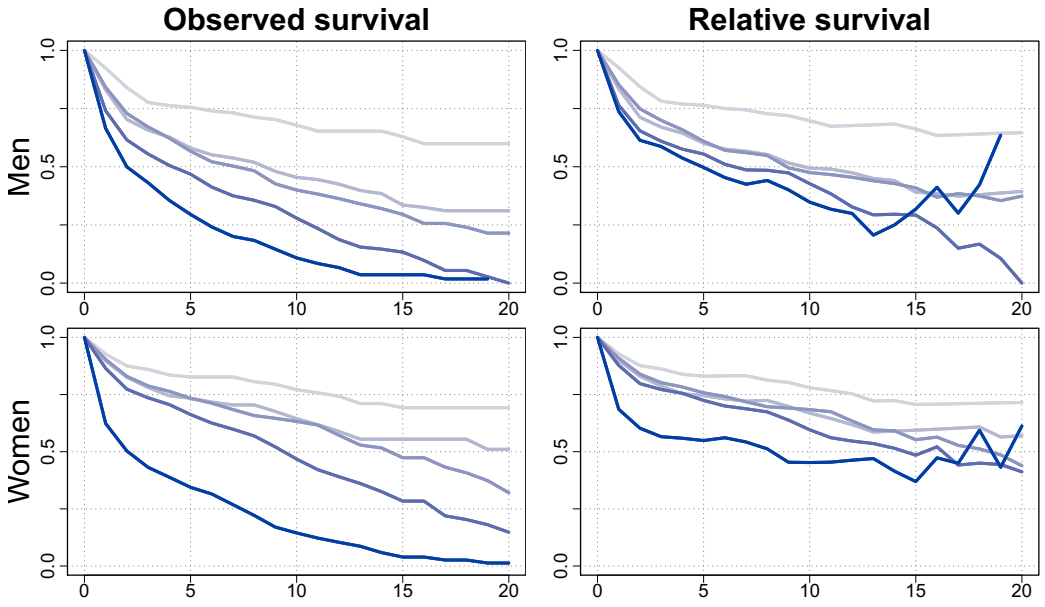

Figure 2: Comparison between the observed survival proportion and the Ederer II estimator of net survival in five age groups. Plots on the left show observed survival, while plots on the right show relative survival. Corresponding colours and age groups go sequentially from the lightest blue referring to the youngest age group to the darkest blue for the oldest age group. The age groups are 0-44 years, 45-54 years, 55-64 years, 65-74 years and 75-100 years.

The five-year survival estimates were similar in the age groups 45-54 years and 55-64 years for both sexes. Otherwise the older age groups had a lower estimated survival. A clear difference between the two estimates, within an age group, was seen only in the two oldest age groups for both sexes.

The Ederer II estimator started behaving illogically at 12 to 15 years after diagnosis in the oldest age group for both sexes. The relative survival curve of the age group 65-74 years for men had a rapid decrease, and a rising spike

Age (years)	Method	Sex			
		Men	(SE)	Women	(SE)
0-44	Observed	76	(3.4)	83	(3.3)
	Relative	76	(3.4)	83	(3.3)
45-54	Observed	58	(2.5)	73	(3.1)
	Relative	60	(2.6)	75	(3.1)
55-64	Observed	57	(2.1)	73	(2.4)
	Relative	61	(2.2)	76	(2.5)
65-74	Observed	47	(2.4)	66	(2.3)
	Relative	55	(2.8)	72	(2.5)
75-100	Observed	30	(2.7)	34	(1.8)
	Relative	50	(4.1)	55	(2.7)

Table 5: The observed survival proportions per 100 and the Ederer II estimators of net survival per 100 at five years after diagnosis, stratified by age (years) at the time of diagnosis.

for women, at around 15 to 16 years after diagnosis. The observed survival was close to zero for patients in the oldest age group for both sexes after 13 to 15 years after diagnosis. For men the observed survival dropped to a very low level also in the age group 65-74 years at around 17 years after diagnosis.

In Table 6, the numbers of patients alive in the two oldest age groups at 10, 15 and 20 years after diagnosis are presented. There were only few observations after 15 years of follow-up. The last man in the oldest age group had a censored survival time after 19 years of follow-up. If there were only few patients left in a particular group and no deaths were observed in an interval, the relative survival started behaving illogically, i.e. jumping up and down.

When stratified by age, the relative survival indicated smaller differences in survival between the age groups, than the observed survival. That is, until about ten years after diagnosis, when the stratified group sizes were still sufficient.

Age (years)	Time	Sex	
		Men	Women
65-74	10 years	68	123
	15 years	14	33
	20 years	1	7
75-100	10 years	20	58
	15 years	3	7
	20 years	0	1

Table 6: Number of patients alive in the two oldest age groups at 10, 15 and 20 years after diagnosis.

4.4 Survival by the stage of cancer at the time of diagnosis

In Figure 3, the observed and the relative survival curves are presented, stratified by the stage of cancer at the time of diagnosis. In Table 7, the five-year survival estimates are presented by the stage of cancer and sex for both observed and relative survival.

The estimated survival dropped below 0.5 only two to three years after diagnosis for both sexes, if the cancer was diagnosed as "Spread". The relative survival curve seemed to be shifted more for the "Local" cancer cases in the later time points than for the "Spread" cancer cases.

Stage	Method	Sex			
		Men	(SE)	Women	(SE)
Local	Observed	67	(1.7)	70	(1.6)
	Relative	77	(1.9)	81	(1.8)
Spread	Observed	36	(1.8)	36	(2.1)
	Relative	41	(1.9)	42	(2.4)
Unknown	Observed	52	(2.7)	59	(2.5)
	Relative	60	(3.1)	72	(2.9)

Table 7: The observed survival proportions per 100 and the Ederer II estimators of net survival per 100 at five years after diagnosis stratified by the stage of cancer at the time of diagnosis.

It was expected that cases of cancer that are diagnosed as "Local" would have a better estimated survival than cases diagnosed as "Spread", and at five years after diagnosis, the difference was clear. Survival for patients with cancer diagnosed with an "Unknown" stage was between survival of cancers diagnosed as "Local" and "Spread".

There was little difference between the survival of men and women, when stratified by the stage of cancer at the time of diagnosis. Especially patients with cancer diagnosed as "Spread" had a similarly low survival, regardless of their sex. The difference between men and women with cancer diagnosed as "Local" was also small.

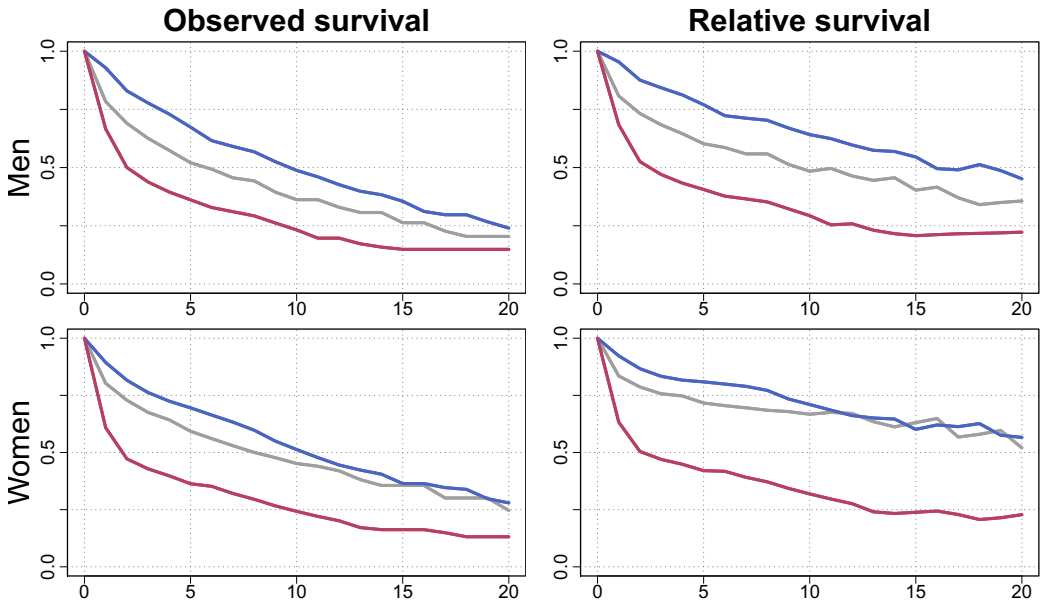

Figure 3: Comparison between the observed survival proportion and the Ederer II estimator of net survival by cancer stage at diagnosis. Plots on the left show observed survival, while plots on the right show relative survival. Corresponding colours and cancer stage: "Local" blue, "Spread" red and "Unknown" gray.

4.5 Survival by the calendar year at the time of diagnosis

In Figure 4, the observed and the relative survival curves are presented for men and women, stratified by the calendar year at the time of diagnosis. Patients were grouped into those diagnosed between 1994-2003, and those diagnosed between 2004-2013. Here survival curves are presented only for the first ten years of follow-up, since that was the maximum follow-up time for the latter group. In Table 8, the five-year and the ten-year survival estimates are presented for both sexes and groups.

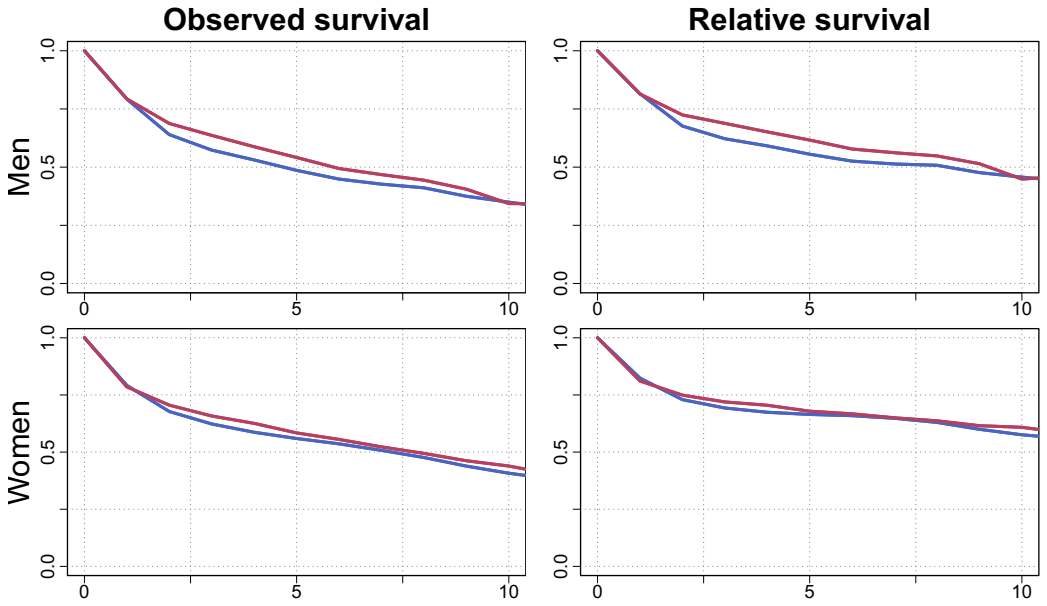

Figure 4: Comparison between observed survival proportion and the Ederer II estimator of net survival by the calendar year of diagnosis. Plots on the left show observed survival, while plots on the right show relative survival. Corresponding colours and calendar years: 1994-2003 blue, 2004-2013 red.

There was little difference in the ten-year estimates within any grouping. Instead some difference was found in the five-year estimates of survival. Accordingly the survival from cancer of the oral cavity had not changed noticeably over time.

Calendar time	Time	Method	Sex			
			Men	(SE)	Women	(SE)
1994-2003	5 years	Observed	49	(1.7)	56	(1.8)
		Relative	56	(2.0)	66	(2.1)
	10 years	Observed	35	(1.7)	41	(1.8)
		Relative	46	(2.1)	58	(2.5)
2004-2013	5 years	Observed	54	(1.6)	58	(1.6)
		Relative	62	(1.7)	68	(1.8)
	10 years	Observed	34	(2.6)	44	(2.2)
		Relative	45	(3.2)	61	(2.9)

Table 8: The observed survival proportions per 100 and the Ederer II estimators of net survival per 100 at five and ten years after diagnosis stratified by the calendar year at the time of diagnosis.

4.6 Comparison of the Pohar Perme and the Ederer II estimators of net survival

In Figure 5, the Ederer II and the Pohar Perme estimators of net survival are compared in three different age groups introduced in Table 2. As mentioned in Section 2.7, these two methods have been recommended by Seppä et al. (2016). In Table 9 the estimates of net survival at five and ten years after diagnosis are presented for both estimators in the same age groups.

The Ederer II and the Pohar Perme estimators appeared to perform similarly up to five years in all three age groups, but with longer follow-up times the Ederer II estimator had a lower standard error than the Pohar Perme estimator. The difference was seen in Figure 5, in the age group 55-64 years when the follow-up time was extended up to 20 years. This difference in the precision of the estimators was clear in the oldest age group already at 10 years after diagnosis, although both estimators tend to become very unstable at the later time points. A similar effect of unstable estimates was already seen in Figure 2, due to a low number of observations in the older age groups at the later follow-up times, as seen in Table 6.

Furthermore, in Figure 6, the age standardised survival curves for the Ederer II and the Pohar Perme estimators are presented. The estimators performed similarly for the whole follow-up period. After ten years of follow-up, the standard error of the Pohar Perme estimator was larger than the standard error of the Ederer II estimator. This is caused by the fact that the Pohar

Perme estimator is affected more by the low numbers of patients alive at the later time points, and in the older age groups. The Ederer II estimator is affected less because it uses the interval specific weights for each individual, while the Pohar Perme estimator is based on the weighted individual-level observations, as shown in Section 2.7.

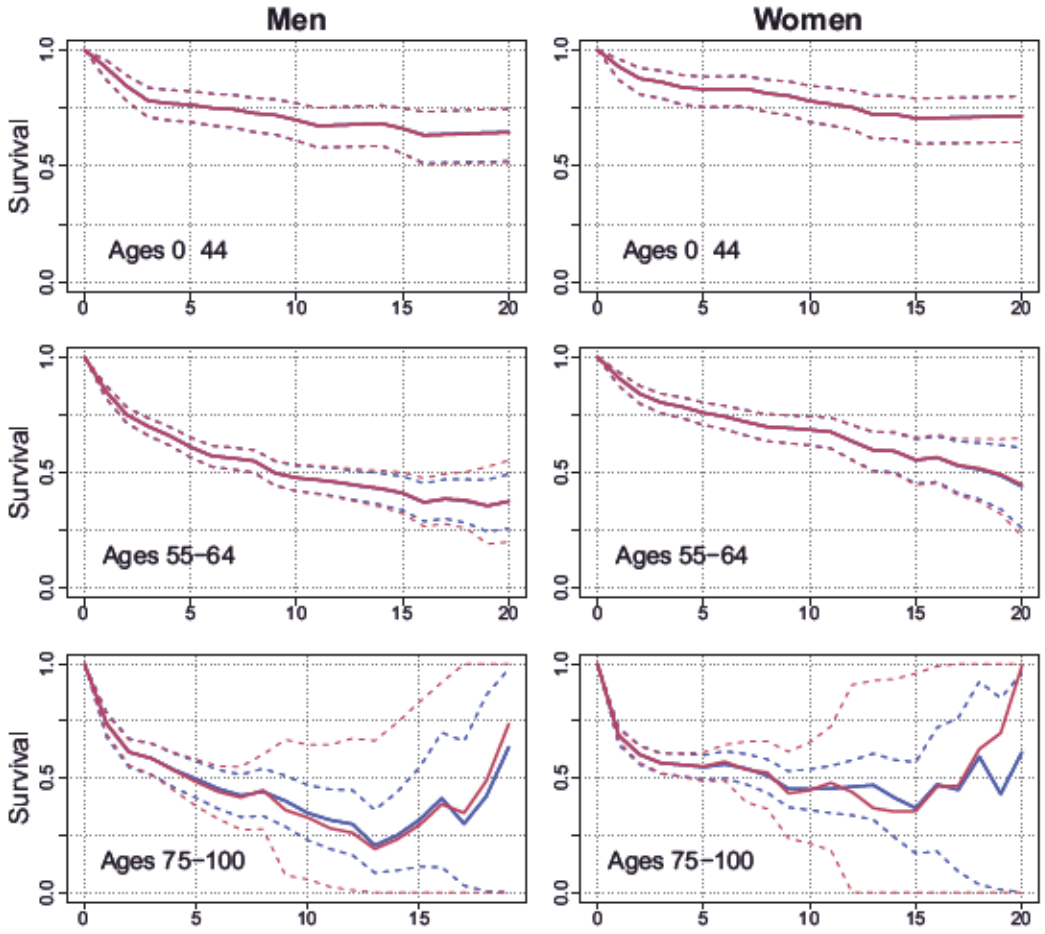

Figure 5: Comparison of the net survival estimators in three age groups. Plots on the left show the survival of men, while plots on the right show the survival of women. The red line corresponds to the Pohar Perme estimator and the blue line corresponds to the Ederer II estimator. The dashed lines show the 95 % confidence intervals.

Both estimators performed similarly, when the five-year survival was estimated. The Ederer II estimator had a lower standard error than the Pohar Perme estimator at the later time points, thus the Ederer II estimator for the net survival probability was used in the descriptive analyses of this thesis.

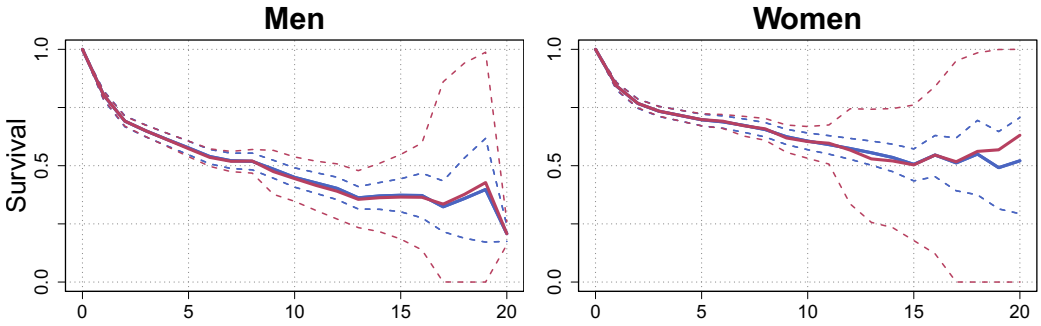

Figure 6: Comparison between the age standardised Ederer II and Pohar Perme estimators. Plot on the left shows the survival of men, while plot on the right shows the survival of women. The red line corresponds to the Pohar Perme estimator and the blue line corresponds to the Ederer II estimator. The dashed lines show the 95 % confidence intervals.

Age group	Time	Sex	Estimator			
			Ederer II	(SE)	Pohar Perme	(SE)
0-44 years	5 years	Men	76	(3.4)	76	(3.3)
		Women	83	(3.3)	83	(3.2)
	10 years	Men	70	(4.1)	70	(4.1)
		Women	78	(1.0)	78	(3.9)
55-64 years	5 years	Men	61	(2.2)	61	(2.2)
		Women	76	(2.5)	76	(2.4)
	10 years	Men	48	(2.8)	48	(3.0)
		Women	68	(3.2)	68	(3.3)
75-100 years	5 years	Men	50	(4.1)	48	(5.3)
		Women	55	(2.7)	55	(3.2)
	10 years	Men	35	(6.3)	33	(17.5)
		Women	45	(4.6)	45	(11.9)

Table 9: The Ederer II and the Pohar Perme estimators of net survival per 100 at five and ten years after diagnosis.

5 Discussion

5.1 The concept of net survival

Net survival is a hypothetical quantity which, in this thesis, was mainly estimated using the method of relative survival. Relative survival is especially suitable when large amounts of data are available, which was the case with the national register data used in this population-based study (Henson and Ries 1995). In the literature, there seems to be some inconsistency in the definition and the interpretation of the survival concepts (Ellis et al. 2014), e.g. some use the term net survival as a synonym for cause-specific survival, while it can be also used like the term net probability in the theory of competing risks (Dickman et al. 2004).

A relative survival ratio is defined as the ratio of the observed survival proportion in the patient group, to the expected survival proportion in a similar group at the beginning of the interval with respect to factors affecting mortality, except the disease under study. It is sometimes interpreted as the proportion of patients alive at the end of an interval to the patients alive at the beginning of the interval, provided that patients are only dying because of the interesting disease. This interpretation gives basis for drawing the relative survival curves. (Hakulinen 1977)

The relative survival ratio and the net survival proportion were thought to be the same quantity until quite recently, with the former serving as an estimate of the latter, although these two can be very different in practice. The relative survival ratio has a clear interpretation in the real world (Pohar Perme, Stare, and Estève 2012), but is often less desirable than the net survival proportion because it is stronly dependent on the population mortality trends. Though the concept of net survival is commonly used, it must be kept in mind that it is not a real world measure. (Rebolj Kodre and Pohar Perme 2013)

As mentioned in Section 2.3 the interpretation of net survival, as the survival probability in the hypothetical situation where patients can only die of cancer, requires strong assumptions about the independency of competing risks (Rebolj Kodre and Pohar Perme 2013). This assumption can not be tested, and generally the assumption can be taken to be invalid. Hence the observed net survival does not have a proper probability interpretation (Seppä 2012), and the interpretation of net survival as a hypothetical probability of dying of cancer is usually not even of interest (Rebolj Kodre and Pohar Perme 2013).

Two different estimators were compared in Section 4.6. Both estimators of net

survival, the Pohar Perme and the Ederer II, performed similarly when a short term net survival was estimated, e.g. five-year survival, and a large quantity of data was available. When a long term net survival was estimated, and less data were available, the Pohar Perme estimator was more prone to error. Even if the Ederer II estimate might be too optimistic about it's accuracy, it was used in the descriptive analyses, because survival up to 20 years after diagnosis was of interest in this thesis. Both estimators also showed a rising survival curve at the later time points, which cannot be interpreted in a meaningful way. Similar results, about the accuracy of these two estimators, were also reported earlier in Seppä, Hakulinen, and Pokhrel (2015) and Seppä et al. (2016).

5.2 Results of the survival analysis

The patients' age and stage of cancer at the time of diagnosis were used to stratify the analyses. These factors clearly affected the survival of the patients heavily. Patient's age and stage of cancer at the time of diagnosis are well known factors that affect the prognosis of cancer (Joensuu et al. 2013).

There was little or no difference found between the observed and the relative survival estimates in the two youngest age groups. A clear difference between the two estimators was only seen in the two oldest age groups. Older patients have a lower expected survival than younger patients, in the general population, which elevates their relative survival compared to observed survival.

The importance of the stage of cancer at the time of diagnosis as a prognostic factor was clearly seen. Patients diagnosed with cancer that had already spread at the time of diagnosis had their estimated survival almost halved, compared to those who were diagnosed with cancer that was local in the oral cavity. There was less difference in the estimated survival between sexes when the analyses were stratified by the stage of cancer.

There were more individuals diagnosed with cancer in the calendar period 2004-2013 than in the previous ten-year calendar period, 1994-2003. The increase in cancer cases might be due to many reasons. The fact that people tend to live longer is mentioned as a primary reason by Finnish Cancer Registry (2020). NORDCAN (Danckert et al. 2020; Engholm et al. 2010) also estimates the number of new cases per year rising annually in the latest ten years by 2.0 % for both sexes, and the number of deaths per year due to cancer of the oral cavity by 2.2 % and 2.6 % for men and women respectively. The relative five-year survival per 100 in 2012-2016 was 60 [55-64] and 71 [67-75]

for men and women respectively. Similarly, rising numbers of oral cancer cases were reported by The SEER Cancer Statistics Review (CSR) (Howlader et al. 2020) and by Weir et al. (2015) in the United States.

The estimated five-year relative survival ratio was similar in both ten-year calendar periods in the analyses of this thesis. The reported relative survival by NORDCAN (Danckert et al. 2020; Engholm et al. 2010) was also similar. Even if the number of cancer cases is rising, the survival from cancer of the oral cavity has stayed similar over time. The SEER CSR reports a slow, but steady, rising trend in the five-year relative survival ratio of oral cancer patients (Howlader et al. 2020). Similar trend in Finland cannot be deduced from the results of this thesis.

Men and women had different distributions of age in the diagnosed cases of cancer of the oral cavity. Also, proportionally more men were diagnosed with cancer that had spread at the time of diagnosis, compared to women. As said before, patient's age and the stage of cancer at the time of diagnosis are important factors for the expected survival from cancer. Overall, men have lower five-year and ten-year survival estimates than women.

References

Carstensen, B., M. Plummer, E. Läärä, and M. Hills. 2019. *Epi: A Package for Statistical Analysis in Epidemiology.* https://CRAN.R-project.org/package=Epi.

Collett, D. 2015. *Modelling Survival Data in Medical Research.* Third edition. Chapman & Hall/CRC Texts in Statistical Science Series. Boca Raton: CRC Press, Taylor & Francis Group.

Danckert, B., J. Ferlay, G. Engholm, H.L. Hansen, T.B. Johannesen, S. Khan, J.E. Køtlum, et al. 2020. "NORDCAN: Cancer Incidence, Mortality, Prevalence and Survival in the Nordic Countries, Version 8.2 (26.03.2019)." Accessed February 3. http://www.ancr.nu.

Dickman, P.W., and E. Coviello. 2015. "Estimating and Modeling Relative Survival." *The Stata Journal* 15 (1): 186–215. doi:10.1177/1536867X1501500112.

Dickman, P.W., A. Sloggett, M. Hills, and T. Hakulinen. 2004. "Regression Models for Relative Survival." *Statistics in Medicine* 23 (1): 51–64. doi:10.1002/sim.1597.

Ellis, L., L.M. Woods, J. Estève, S. Eloranta, M.P. Coleman, and B. Rachet. 2014. "Cancer Incidence, Survival and Mortality: Explaining the Concepts." *International Journal of Cancer* 135 (8): 1774–82. doi:10.1002/ijc.28990.

Engholm, G., J. Ferlay, N. Christensen, F. Bray, M.L. Gjerstorff, A. Klint, J.E. Køtlum, E. Ólafsdóttir, E. Pukkala, and H.H. Storm. 2010. "NORDCAN–a Nordic Tool for Cancer Information, Planning, Quality Control and Research." *Acta Oncologica* 49 (5): 725–36. doi:10.3109/02841861003782017.

Finnish Cancer Registry. 2020. Finnish Cancer Registry website. Accessed February 18. https://cancerregistry.fi/.

Gehan, E.A. 1969. "Estimating Survival Functions from the Life Table." *Journal of Chronic Diseases* 21 (9-10): 629–44. doi:10.1016/0021-9681(69)90035-6.

Hakulinen, T. 1977. "On Long-Term Relative Survival Rates." *Journal of Chronic Diseases* 30 (July): 431–43. doi:10.1016/0021-9681(77)90036-4.

Henson, D.E., and L.A. Ries. 1995. "The Relative Survival Rate." *Cancer* 76 (10): 1687–8. doi:10.1002/1097-0142(19951115)76:10<1687::AID-CNCR2820761002>3.0.CO;2-I.

Howlader, N., A.M. Noone, M. Krapcho, D. Miller, A. Brest, M. Yu, J. Ruhl, et al. 2020. "SEER Cancer Statistics Review, 1975-2016, National Cancer

Institute. Bethesda, Md, Https://Seer.cancer.gov/Csr/1975_2016/, Based on November 2018 Seer Data Submission, Posted to the Seer Web Site, April 2019." Accessed February 3.

Joensuu, H., P-L. Kellokumpu-Lehtinen, S. Jyrkkiö, M. Kouri, and L. Teppo. 2013. *Syöpätaudit [Online].* Helsinki: Kustannus Oy Duodecim. https://www.oppiportti.fi/op/opk04504.

Leung, K-M., R.M. Elashoff, and A.A. Afifi. 1997. "Censoring Issues in Survival Analysis." *Annual Review of Public Health* 18 (1): 83–104. doi:10.1146/annurev.publhealth.18.1.83.

Miettinen, J., and M. Rantanen. 2019. *PopEpi: Functions for Epidemiological Analysis Using Population Data.* https://CRAN.R-project.org/package=popEpi.

Pohar Perme, M., J. Stare, and J. Estève. 2012. "On Estimation in Relative Survival." *Biometrics* 68 (1): 113–20. doi:10.1111/j.1541-0420.2011.01640.x.

R Core Team. 2019. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Rebolj Kodre, A., and M. Pohar Perme. 2013. "Informative Censoring in Relative Survival." *Statistics in Medicine* 32 (27): 4791–4802. doi:10.1002/sim.5877.

RStudio Team. 2016. *RStudio: Integrated Development Environment for R.* Boston, MA: RStudio, Inc. http://www.rstudio.com/.

Seppä, K. 2012. "Quantifying Regional Variation in the Survival of Cancer Patients." PhD thesis, Oulu: University of Oulu. http://urn.fi/urn:isbn:9789526200118.

Seppä, K., T. Hakulinen, and A. Pokhrel. 2015. "Choosing the Net Survival Method for Cancer Survival Estimation." *European Journal of Cancer* 51 (9): 1123–9. doi:10.1016/j.ejca.2013.09.019.

Seppä, K., T. Hakulinen, E. Läärä, and J. Pitkäniemi. 2016. "Comparing Net Survival Estimators of Cancer Patients." *Statistics in Medicine* 35 (11): 1866–79. doi:10.1002/sim.6833.

Weir, H.K., T.D. Thompson, A. Soman, B. Møller, and S. Leadbetter. 2015. "The Past, Present, and Future of Cancer Incidence in the United States: 1975 Through 2020." *Cancer* 121 (11): 1827–37. doi:10.1002/cncr.29258.

World Health Organization, WHO. 2013. *International Classification of Dis-*

eases for Oncology (Icd-O) - 3rd Edition, 1st Revision. Publications. 3rd ed.

Zeileis, A., J.C. Fisher, K. Hornik, R. Ihaka, C.D. McWhite, P. Murrell, R. Stauffer, and C.O. Wilke. 2019. "Colorspace: A Toolbox for Manipulating and Assessing Colors and Palettes." ArXiv 1903.06490. arXiv.org E-Print Archive. http://arxiv.org/abs/1903.06490.

Zeileis, A., K. Hornik, and P. Murrell. 2009. "Escaping Rgbland: Selecting Colors for Statistical Graphics." *Computational Statistics & Data Analysis* 53 (9): 3259–70. doi:10.1016/j.csda.2008.11.033.