University of Tartu
Faculty of Science and Technology
Institute of Mathematics and Statistics

Kai Budrikas

The impact of social inequality on stage at diagnosis of colon and rectal cancers in Denmark

Mathematical Statistics

Bachelor's Thesis (9 ECTS)

Supervisors:
Christian Dehlendorff
Krista Fischer

Tartu 2021

THE IMPACT OF SOCIAL INEQUALITY ON STAGE AT DIAGNOSIS OF COLON AND RECTAL CANCERS IN DENMARK

Bachelor's thesis

Kai Budrikas

Abstract. The aim of this bachelor's thesis is to study whether socioeconomic indicators such as income and education have had an impact on at which stage colon and rectal cancers (ICD-10 C18 and C20) were diagnosed in Denmark from 2004 to 2018. The study is based on data, which was provided by the Danish Cancer Society Research Center and involved information on 58 217 colon and rectal cancer cases from the period of study. The thesis gives an overview of these two cancer types, its changes and trends during the observed time period, introduces the statistical methods used in the analysis and presents the findings with possible reasons for the results obtained from this study.

CERCS research specialisation: P160 statistics, operation research, programming, actuarial mathematics.

Key words: colon cancer, rectal cancer, social inequality, cancer staging, logistic regression.

SOTSIAALSE EBAVÕRDSUSE MÕJU KÄÄR- JA PÄRASOOLEVÄHI DIAGNOOSI STAADIUMILE TAANI ANDMETE PÕHJAL

Bakalaureusetöö

Kai Budrikas

Lühikokkuvõte. Käesoleva bakalaureusetöö eesmärk on uurida, kuidas mõjutasid sotsiaalmajanduslikud faktorid nagu sissetulek ja haridus käär- ja pärasoolevähi (ICD-10 C18 ja C20) diagnoosi staadiumit Taanis aastatel 2004-2018. Uurimus põhineb Taani Vähiseltsilt saadud andmetel, mis hõlmavad endas teavet vaadeldud ajaperioodil diagnoositud 58 217 käär- ja pärasoolevähijuhtumi kohta. Töös antakse ülevaade mõlemast vähitüübist, nende muutustest Taanis aastatel 2004-2018, tutvustatakse kasutatud statistilisi meetodeid ning arutletakse analüüsi tulemuste ja nende võimalike põhjuste üle.

CERCS teaduseriala: P160 statistika, operatsioonanalüüs, programmeerimine, finants- ja kindlustusmatemaatika.

Märksõnad: käärsoolevähk, jämesoolevähk, sotsiaalne ebavõrdsus, vähi staadium, logistiline regressioon.

Contents

Introduction

About 5000 people develop colon or rectal cancer in Denmark annually (Danish Cancer Society, 2021). According to the World Health Organization (2020), colorectal cancer was the most common cancer type in Denmark in 2018 with the second highest mortality level. It is well-known that stage at diagnosis has a huge impact on survival in most cancer types. Early diagnosis is thus crucial, which however to some degree is influenced by the patient's health literacy and health seeking behavior. Studying how socioeconomic factors impact stage at diagnosis gives insight on how these factors also influence the survival of colorectal cancer patients.

The purpose of this bachelor's thesis is to give an overview of how socioeconomic indicators such as education and income have affected colon and rectal cancer's stage at diagnosis distribution from 2004 to 2018 in Denmark. This thesis is a part of a continuous monitoring of social inequality in cancer. The data of four different registries with information on 39 014 colon and 19 203 rectal cancer patients of our period of study was provided by the Danish Cancer Society Research Center.

The thesis is divided into two parts – a theoretical and an empirical part. The aim of the theoretical part is to give insight on colon and rectal cancer overall and on its recent trends to do with social inequalities in Denmark. The part also introduces the statistical methods used in this thesis – logistic regression modeling and splines. An overview of how the data was handled before the beginning of the analysis is also provided.

The second part of the thesis introduces the data used in this study and describes its changes throughout the observed time period by socioeconomic factors. The part also provides an overview of the regression models and results of the analysis done according to the methods presented in the theoretical section. Differences in the stage at diagnosis distributions by educational and disposable income levels are presented as odds and odds ratios from 2004 to 2018.

The author is extremely grateful to Christian Dehlendorff, who accepted her to do her internship/thesis at the Danish Cancer Society Research Center and for providing a lot useful tips for statistical programming and analyses. The author would also like to thank her supervisor Krista Fischer for the advice and feedback given when structuring the thesis and interpreting the results.

1 Background

The aim of this chapter is to give an overview of colon and rectal cancer, its temporal trends in incidence in Denmark and offer explanations for the upcoming results of the thesis. The chapter also discusses the concept of social inequality and its impact on cancer in Denmark according to previous studies.

1.1 Colorectal Cancer

The colon is the longest part of the large intestine of a human. The organ is in charge of absorbing water and any remaining nutrients from partially digested food. The indigestible material, stool, is carried on through the colon, stored in the rectum and then defecated through the anus. (National Cancer Institute, 2018)

Colorectal cancer begins in the colon or the rectum and can be called either colon or rectal cancer, accordingly. The cancer starts when healthy cells in the lining of the colon or rectum begin growing out of control, thus forming a lump called a tumor. If the tumor is malignant, meaning it can grow and spread to other parts of the body, it is cancerous. Otherwise, if the tumor can grow but not spread, it is called a benign tumor. The changes in the cells can be caused by both genetic and environmental factors or simply by random mutations during cell division, i.e. errors during the DNA-copying, and usually take years to develop. (American Society of Clinical Oncology, 2021)

Most colon and rectal cancers are adenocarcinomas, which means the cancer is located in the cells that line the inside tissue of the organ. Other types of colorectal cancers include neuroendocrine tumor of the gastrointestinal tract, gastrointestinal stromal tumor, small cell carcinoma and lymphoma. (American Society of Clinical Oncology, 2021)

Cancer types examined in this thesis and their codes according to the International Statistical Classification of Diseases and Related Health Problems (World Health Organization, 2019) are the following:

- C18, malignant neoplasm of colon;
- C20, malignant neoplasm of rectum.

Colorectal cancer has several risk factors to do with lifestyle, the environment or genetics. Being overweight and/or physically inactive is believed to increase the risk of developing colorectal

cancer. Certain types of diets, smoking and alcohol usage is also linked with the disease. Out of genetic aspects, a personal history of colorectal cancer, polyps or inflammatory bowel disease is also considered as a risk factor. Older age also increases the chances of developing colorectal cancer, which the case for most cancer types. (American Cancer Society, 2020)

The progression of a cancerous tumor can be described by its stage at diagnosis. Stage specifies how large the tumor is and if it has spread to other parts of the body. Knowing the stage is necessary for planning the most suitable treatment and for estimating the chances of survival. The most widely used cancer staging system is the TNM system, where T describes the size and extent of the main tumor, N refers to the number of nearby lymph nodes that have cancer and M determines whether the cancer has metastasized (spread to other parts of the body). (National Cancer Institute, 2015)

According to an American study (O'Connell, Maggard, and Ko, 2004), survival of cancer patients depends greatly on at which stage the disease gets diagnosed. In the same analysis, five-year survival rates are estimated to more than 90% if colon cancer is diagnosed in a localized stage and less than 10% when detected with distant metastasis. Since most patients with colorectal cancer do not show any high-risk symptoms such as severe anemia or rectal bleeding, early-stage diagnoses in symptomatic patients are rare (Hamilton, 2009). Frederiksen et al. (2008) also suggest that rectal cancer tends to have more definite symptoms than colon cancer and this means that a patient has a better chance to react to them and thus seek a physician earlier. Healthcare seeking behaviour is known to be associated with socioeconomic factors (e.g. education).

Studies show that screening can reduce mortality of bowel cancer by 15-30% among people over 50 years of age (Danish Cancer Society, 2021). A nationwide colorectal cancer screening programme was implemented in Denmark in January 2014 and the first invitations were sent out in March 2014 (Larsen et al., 2018). The programme is targeted at individuals aged 50-74 years and they are invited every second year (Danish Cancer Society, 2021).

1.2 Social Inequality in Cancer Diagnosis

Social inequalities have a strong impact on all stages in the cancer trajectory: lifestyle risk factors, diagnosis, stage at diagnosis, treatment, survival and end-of-life care. According to Mackenbach et al. (2008), large differences in disease risk have been observed many times in association with socioeconomic factors, such as household income, educational level and occupation. The study also suggests that these inequalities could be reduced, for example by improving educational

opportunities or by income distribution.

Socioeconomic indicators are also said to be associated with the incidence of colon and rectal cancer, as stated in a Danish study (Egeberg et al., 2008). Based on the analysis of colon and rectal cancers diagnosed in 1994-2003, more cases came from groups with greater social disadvantage and mostly amongst men regarding cohabitational status, dwelling size, housing tenure and affiliation to the job market. The study also showed that short- and long-term relative survival was lower in groups with poorer socioeconomic indicators, such as education, disposable income and cohabitational status, both for colon and rectal cancer.

Stage at diagnosis in relation to socioeconomic status in Denmark has been studied earlier among patients with colorectal adenocarcinoma diagnosed in 1996-2004. According to the study, a decrease in the risk of being diagnosed with late-stage rectal cancer was seen among older rectal cancer patients who had a high income and lived with a partner at an owner-occupied housing. A lower risk was spotted among younger rectal cancer patients, who were in the higher education (more than 12 years) group. Although there was a statistically significant social gradient discovered among rectal cancer patients, no trend was found in colon cancer patients. The study states that there are social inequalities that associate with a higher risk for distant metastasis of rectal, but not colon cancer. This might be caused by the differences of symptoms. (Frederiksen et al., 2008)

1.3 Life in Denmark

Denmark follows a welfare model which provides opportunities for its inhabitants to pursue their happiness disregarding economic, social, cultural or gendered backgrounds. The Danish strategy incorporates equal rights for all its citizens to tax-funded benefits such as free healthcare and education. The welfare model works as a risk-reducing mechanism, as Danes have less to worry about in their daily life than most people. (Wiking, 2016)

The labor market of Denmark is based on flexibility for employers, security for workers and an active labor market policy, and together they run on mutual benefit for all parties involved. The strategy works well for the majority, as it allows the companies to adapt to changes, thus stay in business, and provides a safety net for workers and the unemployed. Although Denmark has one of the highest tax rates in the world, high taxes are turned into collective well-being. Denmark is frequently ranked as one of the happiest countries in the world, as they are said to be not very anxious during their everyday lives. (Wiking, 2016)

2 Methodology

This chapter introduces statistical models and methods used in this thesis to study the distribution of stage at diagnosis. Data handling done before the beginning of the analysis is also described in this chapter.

2.1 Logistic Regression Model

This section is based on James et al. (2013).

The logistic regression model is a statistical model used when the response variable Y is binary, thus it follows the Bernoulli distribution. The model describes the probability of the dependent variable belonging to one of the two groups – when we witness some event (1) or not (0).

Consider a data set with a response variable Y (with possible outcomes 0 or 1) and explanatory variables $X = (X_1, X_2, \dots X_n)$, where $n \in \mathbb{N}$. In logistic regression, the following *logistic function* is used to model the relationship between $p(X) = P(Y = 1|X)$ and X:

$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+\dots+\beta_n X_n}}{1 + e^{\beta_0+\beta_1 X_1+\dots+\beta_n X_n}}. \tag{1}$$

The function always returns outputs which lie between 0 and 1, so it fits well for modelling probabilities.

For comparing $p(X)$ to the probability of its complementary event $1 - p(X)$, *odds* $\frac{p(X)}{1-p(X)}$ are used. Odds close to 0 imply a low probability of $p(X)$ and large odds vice versa. By altering (1), we can see that the odds from the logistic function can be expressed as

$$\frac{p(X)}{1 - p(X)} = e^{\beta_0+\beta_1 X_1+\dots+\beta_n X_n}. \tag{2}$$

To compare two groups, *odds ratio* can be used. For example, if we wish to find the difference between groups where variable X_1 exists ($X_1 = 1$) and where X_1 does not ($X_1 = 0$), we may calculate the ratio of the odds (odds ratio) of two groups according to (2):

$$OR = \frac{e^{\beta_0+\beta_1(X_1=1)+\beta_2 X_2+\dots+\beta_n X_n}}{e^{\beta_0+\beta_1(X_1=0)+\beta_2 X_2+\dots+\beta_n X_n}} = \frac{e^{\beta_0+\beta_1+\beta_2 X_2+\dots+\beta_n X_n}}{e^{\beta_0+\beta_2 X_2+\dots+\beta_n X_n}} = e^{\beta_1}.$$

So if we were to compare group $X_1 = 1$ to $X_1 = 0$, the odds for $Y = 1$ are e^{β_1} times larger. In this case, the group with $X_1 = 0$ is also called the *reference group*.

By taking the logarithm of both sides of (2), we can also find that

$$\text{logit}(p(X)) := \log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + \ldots + \beta_n X_n. \tag{3}$$

We have now derived the *log-odds* or *logit* of the logistic regression model. By transforming the odds this way, it is now seen that the logit is linear in X.

According to (3), in the logistic regression model, increasing one explanatory variable, for example X_1, by one unit changes the log odds by β_1 (if the other predictor variables stay the same). Similarly, the increase by one unit multiplies the odds by e^{β_1} (2). Since there is no straight-line relationship between $p(X)$ and X, β_1 is not the change of $p(X)$ when X_1 increases by one unit. The rate of change in $p(X)$ per unit depends on the present value of X, but if β_1 is positive then increasing X_1 increases $p(X)$ and if β_1 is negative then increasing X_1 decreases $p(X)$.

Since the model implies that $Y_i \sim \text{Be}(p(X_i))$, the coefficients $\beta_0, \beta_1, \ldots, \beta_n$ in (1) are estimated based on data by using the following *likelihood function*:

$$l\left(\beta_0, \ldots, \beta_n\right) = \prod_{i=1}^{k} p(x_i)^{y_i}(1-p(x_i))^{1-y_i}, \tag{4}$$

where k is the number of observations in the dataset. To estimate the parameters of the model $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_n$, the likelihood function (4) is maximized with respect to the parameters.

2.2 Regression Splines

Regression splines are an extension of polynomials and step functions. For a continuous variable X, this involves dividing the range of X into K distinct regions, where a polynomial function is fit to the data in each region. For keeping the combined function of these subsections smooth and continuous at the region boundaries called *knots*, we also impose three constraints - continuity, continuity of the first derivative and continuity of the second derivative. So if the interval of X is divided into enough regions, this can produce a remarkably flexible fit. (James et al., 2013)

According to Gauthier, Wu, and Gooley (2020), if we were to add an additional boundary constraint that the function is required to be linear before the first knot and after the last knot, we get the *restricted cubic spline*. The number of knots can be chosen freely, but the use of a large or a small number of knots may lead to over- or underfitting the model. An option for choosing the knots would be according to quantiles, as it is done with function rcs of package rms in R (Harrell, 2021).

Suppose we have a continuous variable X. A restricted cubic spline with K knots is modelled as

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 h(x_i, \xi_1) + \ldots + \beta_{K+1} h(x_i, \xi_K) + \epsilon_i =: \beta_0 + \text{rcs}(x_i) + \epsilon_i, \tag{5}$$

where ϵ_i is the *error term* and $h(x, \xi)$ is a specific function called a *truncated power basis function.* The truncated power basis function is defined as

$$h(x, \xi) = \begin{cases} (x - \xi)^3 & \text{if } x > \xi \\ 0 & \text{otherwise,} \end{cases}$$

where ξ is one of the knots. (James et al., 2013; Gauthier, Wu, and Gooley, 2020)

So if we wish to fit a restricted cubic spline according to (5) for some data with K fixed knots, we execute *least squares regression* with an intercept and $K + 1$ predictors of the form X, $h(X, \xi_1)$, $\ldots$, $h(X, \xi_K)$. The method of least squares is a standard model fitting method where the sum of the squares of the *residuals* (the difference between the ith observed response value and the ith response value predicted by the model) is minimized. Fitting a restricted cubic spline with K knots uses $K + 2$ degrees of freedom as we have to estimate $K + 2$ regression coefficients, as shown in (5). (James et al., 2013)

2.3 Data Handling

Descriptive and inferential analyses of this thesis were carried out with R. In order to study all our variables of interest, four separate datasets were used:

- The Danish Cancer Registry with cancer cases and their TNM-classifications;
- Population Registry (from Statistics Denmark) with birthdates and sexes;
- Income Registry (from Statistics Denmark) with disposable incomes;
- Education Registry (from Statistics Denmark) with highest achieved educations.

The crude datasets contained millions of observations and were combined according to a common pseudonymized id variable. Cases analyzed in this study were the first ever cancer diagnoses (except for non-melanoma skin cancer) of people aged 18 or above from 2004 to 2018. There were a total of 39 014 colon and 19 203 rectal cancer cases diagnosed during our period of study. Socioeconomic status was measured by disposable income and highest achieved education, which

were both assessed the year prior to diagnosis to avoid issues with changes in income caused by early cancer symptoms.

Stage of cancer was defined by the so-called TNM-classification and classified into three groups: localized, non-localized and unknown/unclassified. Cancer cases were considered as non-localized if cancer had spread to nearby lymph nodes or if it had metastasized to other parts of the body. Otherwise, if there was no sign that the cancer had spread, it was categorized as localized.

Data on annual disposable income in the Income Registry was coded as age and sex standardized quintiles. The population was split into four groups accordingly: lowest 20%, medium 60%, highest 20% and unknowns. As the level of household disposable income inequality in Denmark measured by the Gini coefficient is one of the lowest across the OECD and wealth is rather equally distributed among the population (Causa et al., 2016), grouping people into quantiles this way shows greater effect when comparing the lower and upper categories to each other.

Information on highest achieved education in the Education Registry was coded with a common AFSP1E variable described by Danmarks Statistik (2015). Education was categorized into four groups accordingly: basic education, vocational or short further education (SFE), medium or long further education (LFE), and unknowns.

3 Descriptive Overview

This chapter introduces the trends of stage at diagnosis of both colon and rectal cancers from 2004 to 2018. The chapter describes the incidences and stage distributions by educational and disposable income groups.

3.1 Colon Cancer

The incidence of colon cancer by educational groups and stage at diagnosis from 2004 to 2018 has been summarised into Table 1. The table shows three educational groups split into three possible categories of stage at diagnosis - non-localized, localized and unknown (NA). As seen from the table, the number of cancer cases remained more or less the same at the start of our study period, but increased in the second half. This is probably caused by the national colorectal cancer screening programme, which began in 2014. The increase in the number of cancer cases with localized stage seems to be the most notable amongst people with medium or long further education, as the number has tripled during the observation period.

Table 1: Colon cancer incidence in Denmark from 2004 to 2018 by education and stage at diagnosis.

	Basic			**Vocational or SFE**			**Medium or LFE**		
Year of diagnosis	**Non-loc**	**Loc**	**NA**	**Non-loc**	**Loc**	**NA**	**Non-loc**	**Loc**	**NA**
2004	508	314	108	340	242	63	131	81	22
2005	503	289	133	354	220	88	166	76	33
2006	471	342	141	405	256	103	152	113	33
2007	520	320	158	431	254	99	169	93	32
2008	559	343	177	471	242	135	182	95	47
2009	557	311	188	454	274	114	190	121	51
2010	545	377	161	478	318	118	227	118	43
2011	548	322	214	534	323	156	211	136	51
2012	519	394	240	474	318	191	216	140	89
2013	491	348	282	465	371	271	212	135	104
2014	518	502	301	536	520	268	207	227	137
2015	514	489	383	502	524	381	227	254	152
2016	545	542	275	566	547	286	276	240	113
2017	559	478	227	625	595	268	293	282	120
2018	553	424	215	613	501	213	277	227	98

To get an insight on how the distribution of stages has changed throughout the years, we can find the proportion of localized stage cancer cases out of all cases with a determined stage at diagnosis, as shown in Table 2. The table implies that the situation has improved in all

educational groups, as the percentage of cancer cases discovered at an earlier stage has risen, most probably due to the screening programme. The table also suggests that the changes were most pronounced amongst those with the longest education.

Table 2: Crude proportions of stage localized diagnoses of colon cancer by education and year of diagnosis. Observations with unknown stages have been excluded.

Year of diagnosis	**Basic**	**Vocational or SFE**	**Medium or LFE**
2004	38%	42%	38%
2005	36%	38%	31%
2006	42%	39%	43%
2007	38%	37%	35%
2008	38%	34%	34%
2009	36%	38%	39%
2010	41%	40%	34%
2011	37%	38%	39%
2012	43%	40%	39%
2013	41%	44%	39%
2014	49%	49%	52%
2015	49%	51%	53%
2016	50%	49%	47%
2017	46%	49%	49%
2018	43%	45%	45%

As Table 3 suggests, the change in disposable income groups has been similar as in educational levels, as the number of cancer cases seems to be on a rise. After the beginning of the screening programme, the number of localized cancer cases has even exceeded the number of non-localized cases at least once in all income groups. There do not seem to be large differences in the colon cancer incidence changes when we compare the group with the lowest income to the one with the highest by stages.

Table 3: Colon cancer incidence in Denmark from 2004 to 2018 by disposable income groups and stage at diagnosis.

	Lowest 20%			Medium 60%			Highest 20%		
Year of diagnosis	**Non-loc**	**Loc**	**NA**	**Non-loc**	**Loc**	**NA**	**Non-loc**	**Loc**	**NA**
2004	230	162	57	661	454	169	234	146	49
2005	228	134	72	696	413	196	242	143	74
2006	236	153	86	692	504	208	213	159	68
2007	246	136	72	737	455	224	234	148	75
2008	269	168	76	778	442	255	250	129	84
2009	256	147	108	760	437	242	252	166	73
2010	268	182	55	810	531	242	251	159	77
2011	262	167	107	835	502	277	243	149	70
2012	236	168	113	794	570	353	217	147	92
2013	229	177	131	764	550	444	221	163	113
2014	247	264	132	800	782	490	249	233	123
2015	244	285	163	800	760	622	237	250	179
2016	237	216	108	885	888	453	306	266	139
2017	254	247	95	922	859	415	332	279	124
2018	244	217	81	924	733	360	312	240	104

Throughout the years of study, the proportion of stage localized colon cancer cases has mostly been below 50% in all disposable income groups, as Table 4 suggests. The percentage has exceeded 50% only after the start of the colorectal screening programme in Denmark. The relative changes between three groups do not seem to be considerably different.

Table 4: Crude proportions of localized stage diagnoses of colon cancer by disposable income and year of diagnosis. Observations with unknown stages have been excluded.

Year of diagnosis	**Lowest 20%**	**Medium 60%**	**Highest 20%**
2004	41%	41%	38%
2005	37%	37%	37%
2006	39%	42%	43%
2007	36%	38%	39%
2008	38%	36%	34%
2009	36%	37%	40%
2010	40%	40%	39%
2011	39%	38%	38%
2012	42%	42%	40%
2013	44%	42%	42%
2014	52%	49%	48%
2015	54%	49%	51%
2016	48%	50%	47%
2017	49%	48%	46%
2018	47%	44%	43%

To conclude, the incidence of colon cancer by educational and disposable income groups seems

to have increased at the end of our observed period of time. While the proportion of cases diagnosed at stage localized also improved around that time, there does not seem to be much difference between the changes when comparing educational or disposable income groups to each other. The changes were most probably caused by the colorectal cancer screening programme introduced in 2014.

3.2 Rectal Cancer

Rectal cancer incidence from 2004 to 2018 by educational groups is shown in Table 5, which suggests that the number of rectal cancers diagnosed with stage localized increased steadily in all educational groups. The table also shows that the incidence of non-localized rectal cancer cases has risen amongst people with vocational, short further, medium or long further education, but decreased only in the group with basic education. The smallest number of rectal cancer cases was in the group with medium or long further education.

Table 5: Rectal cancer incidence in Denmark from 2004 to 2018 by education and stage at diagnosis.

	Basic			**Vocational or SFE**			**Medium or LFE**		
Year of diagnosis	**Non-loc**	**Loc**	**NA**	**Non-loc**	**Loc**	**NA**	**Non-loc**	**Loc**	**NA**
2004	259	145	90	191	126	54	69	40	12
2005	238	142	80	167	126	62	71	44	26
2006	278	156	99	244	141	75	86	56	17
2007	280	133	91	230	117	65	99	46	27
2008	258	141	109	243	149	84	91	57	26
2009	309	144	87	269	154	55	94	43	25
2010	283	159	97	255	152	60	111	55	20
2011	263	166	102	273	163	85	110	52	29
2012	268	139	130	281	179	107	99	75	38
2013	224	152	180	205	201	151	85	66	55
2014	241	170	177	280	242	197	112	93	81
2015	201	195	192	249	258	235	98	91	92
2016	212	195	129	240	248	136	111	114	55
2017	241	208	104	293	263	119	116	135	42
2018	214	171	104	267	288	132	112	117	43

The proportion of rectal cancers being diagnosed in localized stage has increased during our study period, as seen in Table 6. The proportion of these cases has exceeded 50% amongst people with vocational, short further, medium or long further education, but it has been only slightly lower in the group with basic education. The highest proportions were seen after the introduction of the colorectal cancer screening programme. The differences between educational

levels do not seem to be substantial.

Table 6: Crude proportions of stage localized diagnoses of rectal cancer by education and year of diagnosis. Observations with unknown stages have been excluded.

Year of diagnosis	Basic	Vocational or SFE	Medium or LFE
2004	36%	40%	37%
2005	37%	43%	38%
2006	36%	37%	39%
2007	32%	34%	32%
2008	35%	38%	39%
2009	32%	36%	31%
2010	36%	37%	33%
2011	39%	37%	32%
2012	34%	39%	43%
2013	40%	50%	44%
2014	41%	46%	45%
2015	49%	51%	48%
2016	48%	51%	51%
2017	46%	47%	54%
2018	44%	52%	51%

The number of people diagnosed with rectal cancer at an earlier stage seemed to be increasing during the period of our study irrespective of income group, as shown in Table 7. The differences in incidence between the lowest and highest income groups do not seem immense.

Table 7: Rectal cancer incidence in Denmark from 2004 to 2018 by disposable income groups and stage at diagnosis.

	Lowest 20%			Medium 60%			Highest 20%		
Year of diagnosis	Non-loc	Loc	NA	Non-loc	Loc	NA	Non-loc	Loc	NA
2004	113	77	47	336	212	141	120	68	41
2005	112	72	54	308	203	136	104	68	43
2006	142	83	61	385	232	130	120	64	45
2007	137	66	42	393	200	127	112	55	56
2008	126	77	56	395	224	148	110	63	48
2009	150	72	41	408	204	126	135	80	31
2010	131	79	51	418	235	112	131	65	31
2011	141	79	43	400	245	153	128	65	41
2012	154	78	72	423	261	184	95	62	39
2013	112	62	81	315	283	239	101	86	77
2014	112	84	96	434	343	285	113	86	89
2015	108	103	109	353	363	344	102	92	87
2016	85	75	66	380	371	197	110	127	67
2017	116	103	57	409	400	167	139	119	45
2018	98	100	53	409	351	190	99	136	43

As Table 8 suggests, the proportion of rectal cancers being diagnosed in localized stage appeared

to be quite stable from 2004 to 2013 with some minor fluctuation, but then increased steadily in all disposable income levels from 2014 to 2018. The trend in the proportion of cases in localized stage seems be similar between income levels.

Table 8: Crude proportions of localized stage diagnoses of rectal cancer by disposable income and year of diagnosis. Observations with unknown stages have been excluded.

Year of diagnosis	**Lowest 20%**	**Medium 60%**	**Highest 20%**
2004	41%	39%	36%
2005	39%	40%	40%
2006	37%	38%	35%
2007	33%	34%	33%
2008	38%	36%	36%
2009	32%	33%	37%
2010	38%	36%	33%
2011	36%	38%	34%
2012	34%	38%	39%
2013	36%	47%	46%
2014	43%	44%	43%
2015	49%	51%	47%
2016	47%	49%	54%
2017	47%	49%	46%
2018	51%	46%	58%

In conclusion, the data suggests that the proportion of people being diagnosed with rectal cancer at an earlier stage is increasing. The increase began around 2013 and most likely due to the national screening programme, which was introduced at that time. The trends seem to be similar in all educational and disposable income groups.

If we were to compare the two cancer types to each other, it is clear that the incidence of colon cancer is much greater than it is of rectal cancer. The changes for both types of cancer follow a similar path, as the proportions of localized cases remain quite stable at the beginning and middle of our period of observation, but begin to grow somewhere around 2014 when the screening programme started. The differences of educational and disposable income groups for both cancer types do not seem immense, but require a more meticulous analysis. In order to examine our data more thoroughly, we may now apply the regression methods described previously.

4 Analysis

The aim of this chapter is to present the results of the analysis done according to the methodology introduced in Chapter 2. Models were fit with educational and disposable income groups separately, as they are somewhat correlated.

To analyse how the odds of being diagnosed with stage localized varies in different educational and income groups among colon and rectal cancer patients, two kinds of logistic regression models were created – 1) with year of diagnosis and age included linearly in the models and 2) with year of diagnosis and age included in the models through restricted cubic splines.

The log odds of the first type of models with educational levels follow a structure of

$$\text{logit}(p(\text{localized})) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5, \tag{6}$$

where X_1=year, $X_2 = 1$ if sex is female (0 otherwise), $X_3 = 1$ if education is vocational or SFE (0 otherwise), $X_4 = 1$ if education is medium or LFE (0 otherwise) and X_5=age/5. The structure of the models with disposable income remains the same, but $X_3 = 1$ if disposable income is the medium 60% group (0 otherwise) and $X_4 = 1$ if it is the highest 20% group (0 otherwise).

To compare the odds of different groups, odds ratios with 95% confidence intervals were calculated. The reference groups included men with either basic education or who belonged to the lowest income group (depending on the independent variables of the specific model). Year of diagnosis and age at the odds ratios represent the OR for a one and 5-year increase, respectively. Forest plots were composed to illustrate the odds ratios graphically.

To fit a more flexible model, the second kind of modelling involved restricted cubic splines and an interaction between calendar year and education or income. Sex was excluded from these as it didn't show great effect in the first type of models. Log odds of the models with education were found as

$$\text{logit}(p(\text{localized})) = \beta_0 + \text{rcs}(X_1) + \beta_1 X_2 + X_2 \cdot \text{rcs}(X_4) + \beta_2 X_3 + X_3 \cdot \text{rcs}(X_4) + \text{rcs}(X_4), \tag{7}$$

where rcs(X) is the function defined in (5), X_1=age, $X_2 = 1$ if education is vocational or SFE (0 otherwise), $X_3 = 1$ if education is medium or LFE (0 otherwise) and X_4=year. In the models with disposable income, $X_2 = 1$ if disposable income is the medium 60% group (0 otherwise) and $X_3 = 1$ if it is the highest 20% group (0 otherwise).

To compare our educational or income levels, the outcomes of these models were presented as

odds ratios depending on the calendar year, where the reference group had either basic education or belonged to the lowest disposable income class. The odds ratios and log odds from (7) were illustrated graphically by each calendar year.

4.1 Colon Cancer

First, a simple model (6) with educational levels was fitted on colon cancer data and odds ratios for each variable were calculated. As Figure 1 suggests, year and patient's age played an important role in colon cancer diagnosis. The odds of being diagnosed at an early stage if the person was deemed to get colon cancer got slightly better as the year of diagnosis and age increased. Sex did not show any statistical significance in colon cancer's stage at diagnosis and neither did educational levels, as 1 was included in their 95% confidence intervals.

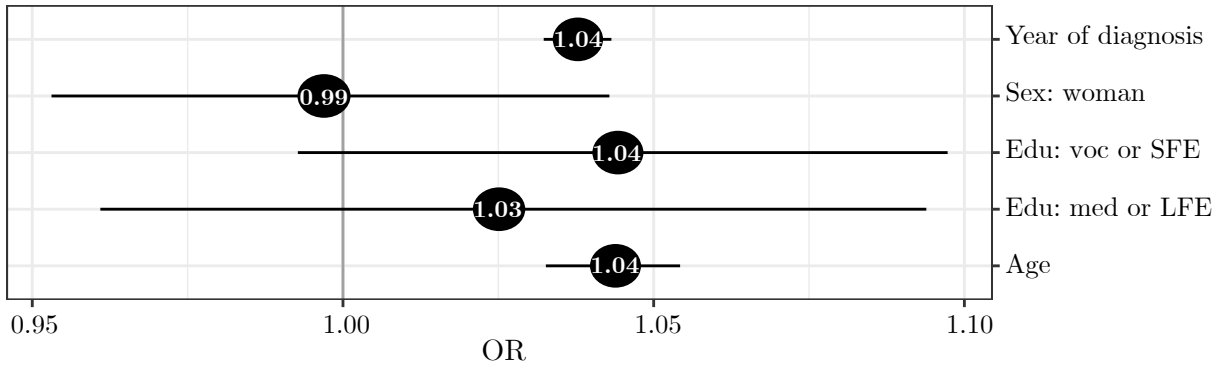

Figure 1: Odds ratio (OR) estimates of colon cancer being diagnosed in stage localized with 95% confidence intervals. Reference group includes men with basic education, who are 5 years younger and who were diagnosed with colon cancer a year before.

A model with educational levels was also fitted according to (7). As Figure 2 shows, educational groups tended to behave differently at the beginning of our study period. Belonging to group vocational or SFE seemed to increase and belonging to group medium or LFE to decrease the odds of being diagnosed at an earlier stage a little when comparing them to the lowest level, basic education. Since all of the confidence intervals are quite wide, we cannot draw any firm conclusions from this model.

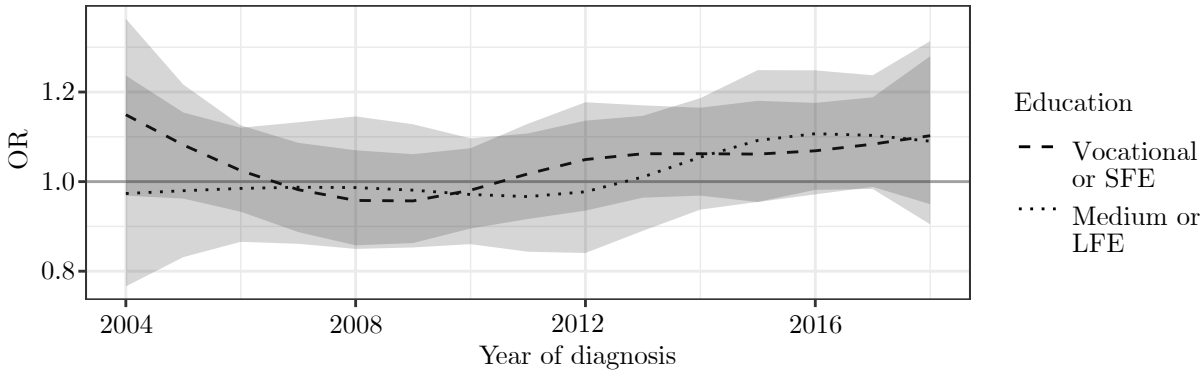

Figure 2: Odds ratio estimates of colon cancer being diagnosed in stage localized from 2004 to 2018 with 95% confidence intervals. Reference group includes people with basic education.

If we look at the log odds estimates of colon cancer being diagnosed at an early stage (Figure 3) calculated from model (7) for people aged 71.9, which is the mean age of our observed colon cancer patients, we also see that the difference between educational levels is not immense, as the confidence intervals overlap. The log odds of early-stage colon cancer were negative from 2004 to about 2013, which means that the probability for being diagnosed at localized stage was lower than 50% at that time. This was also seen from Table 2, as the first time the proportion of localized cases rose over 50% was in 2014 (group medium or LFE), but here in the model we have also adjusted for age. The probability of being diagnosed at an early stage increased gradually for all educational levels from 2011 to 2015, most probably due to the national screening programme, but then started to decrease. The probability of early-stage exceeded the probability of late-stage colon cancer from 2015 to 2016 significantly in groups with vocational or short further and medium or long further education.

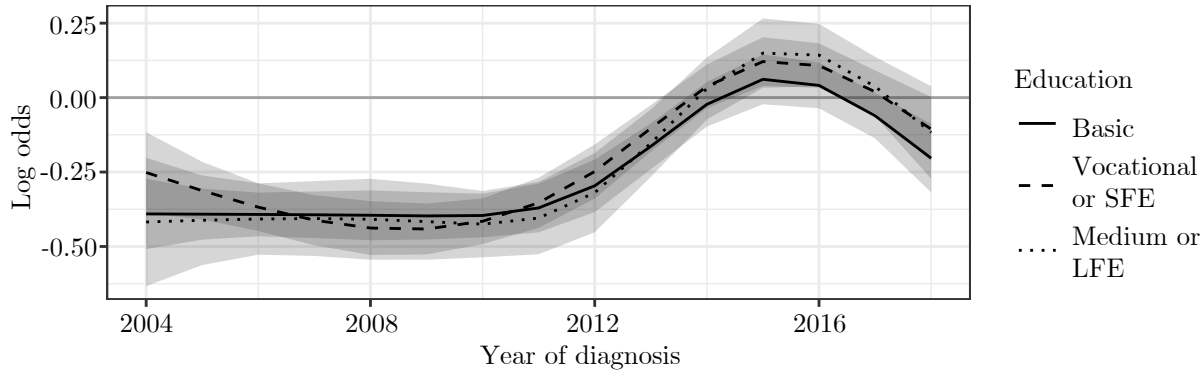

Figure 3: Log odds estimates of colon cancer being diagnosed in stage localized by education for people aged 71.9 years from 2004 to 2018 with 95% confidence intervals.

To study whether disposable income had an effect on stage at diagnosis of colon cancer, a model according to (6) was also fitted for income groups. As Figure 4 suggests, if we were to compare a

person from the lowest income class to someone from the medium or highest income group (and other characteristics remain the same), the first person would have a slightly higher chance of being diagnosed at an earlier stage, although not statistically significantly. Similarly as before with the model with educational levels, the difference according to sex was very small here as well. Later year of diagnosis and older age slightly increased the odds of being diagnosed at an earlier stage with colon cancer, as seen from this model.

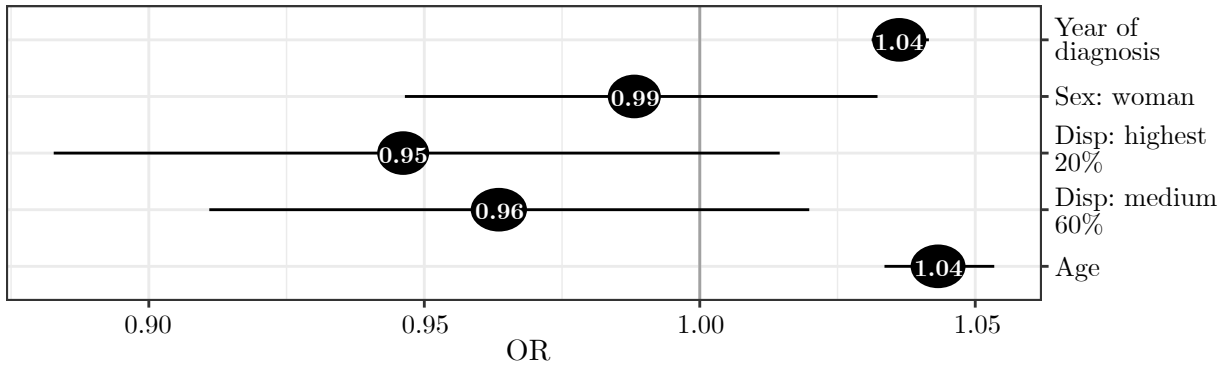

Figure 4: Odds ratio estimates of colon cancer being diagnosed in stage localized with 95% confidence intervals. Reference group includes men from the poorest income group, who are 5 years younger and who were diagnosed with colon cancer a year before.

As implied by the previous model and a model fitted with restricted cubic splines (Figure 5), the odds of early-stage colon cancer behaved rather similarly in all disposable income groups, but with a tendency to a difference between the lowest 20% and the highest 80% in the last part of the observed period. This implies that at that time, the richest people had slightly lower odds for being diagnosed at an earlier stage with colon cancer than the poorest, although not statistically significantly.

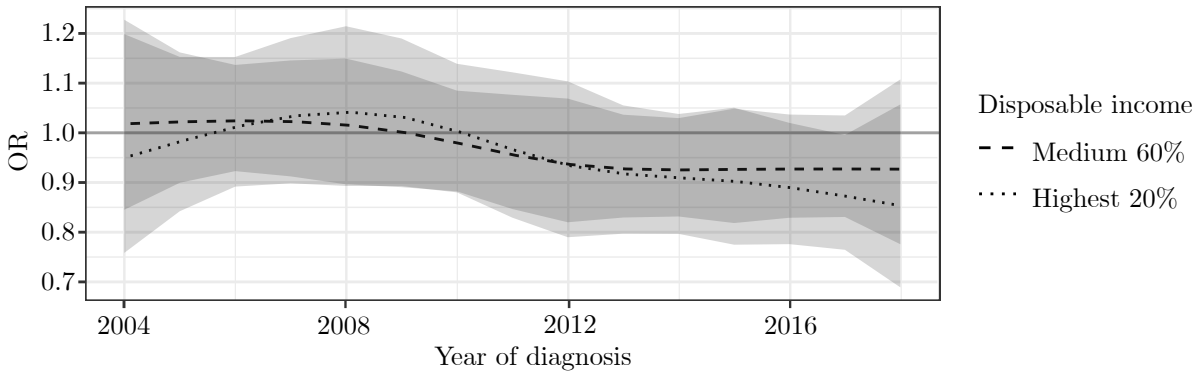

Figure 5: Odds ratio estimates of colon cancer being diagnosed in stage localized from 2004 to 2018 with 95% confidence intervals. Reference group includes people from the poorest income group.

When estimating the log odds of all disposable income levels as shown in Figure 6, we see similar results to what we already witnessed when comparing educational levels (Figure 3). The probability of being diagnosed with an early-stage case of colon cancer has mostly been smaller than the probability of non-localized colon cancer during our observed period of time. Although the odds bettered from 2010 to 2015 in all disposable income levels, the probability of being diagnosed with early-stage colon cancer seemed to be declining during the last years of our study period.

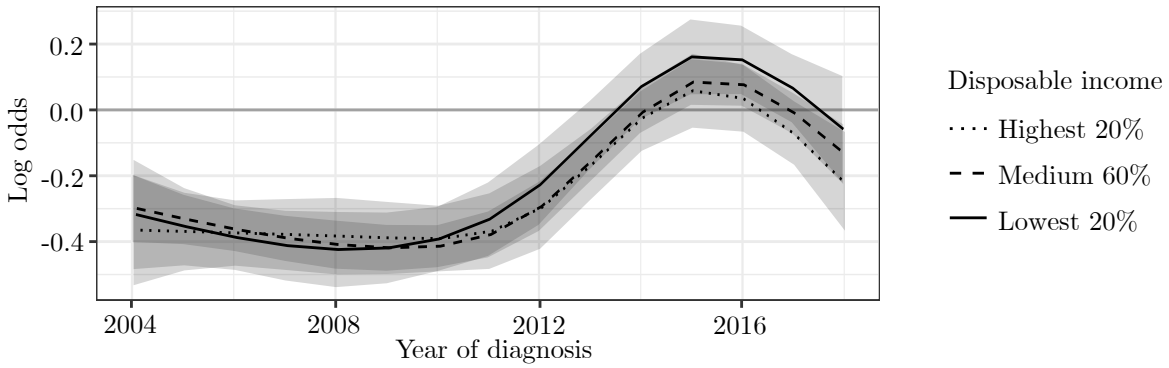

Figure 6: Log odds estimates of colon cancer being diagnosed in stage localized by disposable income for people aged 71.9 years from 2004 to 2018 with 95% confidence intervals.

To conclude, neither educational nor disposable income levels had much effect on at which stage colon cancer was diagnosed during our period of study. These results were as expected, since a similar study in Denmark on data from 1996-2004 stated that the same socioeconomic indicators did not have an major effect on staging of colon cancer at that time, too (Frederiksen et al., 2008). On the other hand, Figures 3 and 6 indicate that the situation has gotten better with time, as the odds of colon cancer being diagnosed in localized stage have improved for those diagnosed later.

4.2 Rectal Cancer

The effects of year of diagnosis, sex, education and age in cancer staging were also studied amongst rectal cancer cases similarly to what was done for colon cancer. Figure 7 suggests, that the odds of being diagnosed with early-stage rectal cancer are 1.19 times higher for a person with vocational or short further education and 1.16 times higher for someone with medium or long further education compared to a person with only basic education. Sex did not prove to be significant here, but year of diagnosis and age both did. Increasing either calendar year by one or age by five years increases the odds of being diagnosed at an early stage 1.05 times.

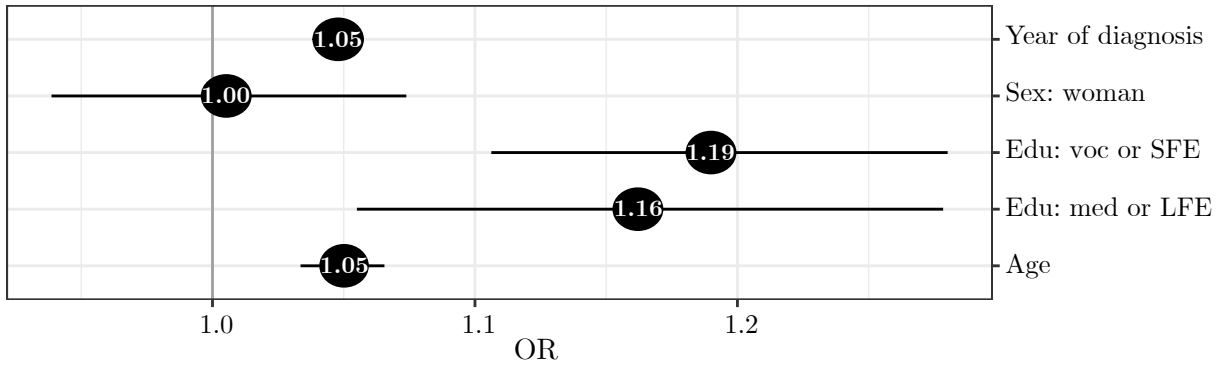

Figure 7: Odds ratio estimates of rectal cancer being diagnosed in stage localized with 95% confidence intervals. Reference group includes men with basic education, who are 5 years younger and who were diagnosed with rectal cancer a year before.

If we were to compare the odds ratios which are calculated according to a model fit with restricted cubic splines, we see that the odds ratio between vocational or short further and basic education proves to be statistically significant (above 1) mostly in the latter part of the period (Figure 8). Thus, we may state that around those times the odds of being diagnosed with rectal cancer at an earlier stage for a person with vocational or short further education were slightly higher than they were for someone with only basic education. A similar conclusion can be drawn for a person with medium or long further education from 2016 to 2018. Generally, the odds ratios of the two highest educational groups seem quite stable or perhaps with a tendency to an increase at the end of our observed period.

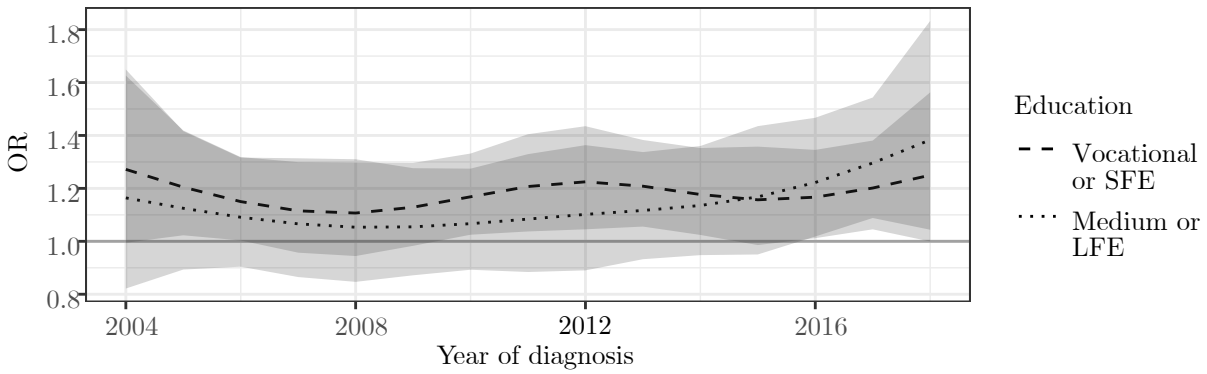

Figure 8: Odds ratio estimates of rectal cancer being diagnosed in stage localized from 2004 to 2018 with 95% confidence intervals. Reference group includes people with basic education.

When estimating the log odds of being diagnosed with early-stage rectal cancer according to the model with restricted cubic splines for the mean age of our observed rectal cancer patients, we see that the probability of being diagnosed with rectal cancer at stage localized has improved during our period of study (Figure 9). This is likely due to the national screening programme

implemented in 2014. Although the differences between the educational levels are not statistically significant, the odds of the group with only basic education have consistently been lower than the odds of the other educational groups from 2004 to 2018.

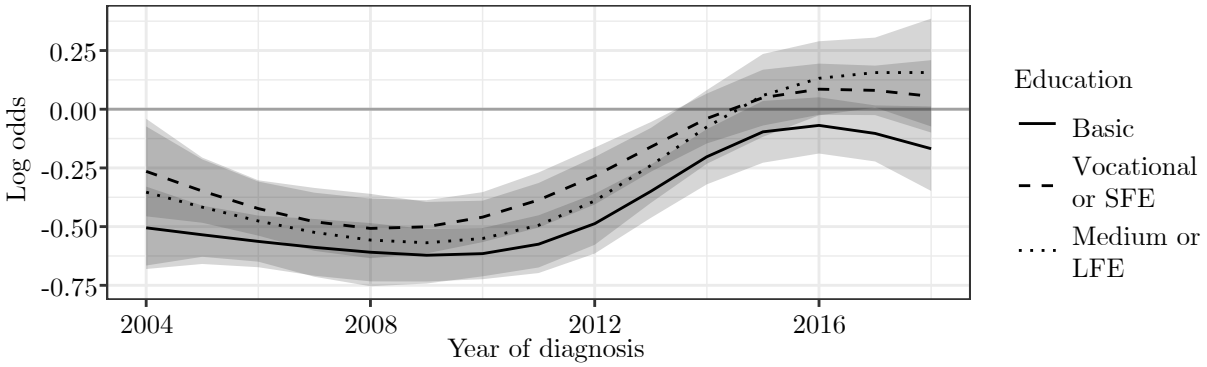

Figure 9: Log odds estimates of rectal cancer being diagnosed in stage localized by education for people aged 69.5 years from 2004 to 2018 with 95% confidence intervals.

As we can see from Figure 10 and as was stated earlier according to the model with educational levels, the odds of localized stage rectal cancer increased with year of diagnosis and age. Sex cannot be claimed as a significant factor this time, too. A higher disposable income might slightly increase the odds of early-stage rectal cancer, but it cannot be concluded on a 0.05 significance level.

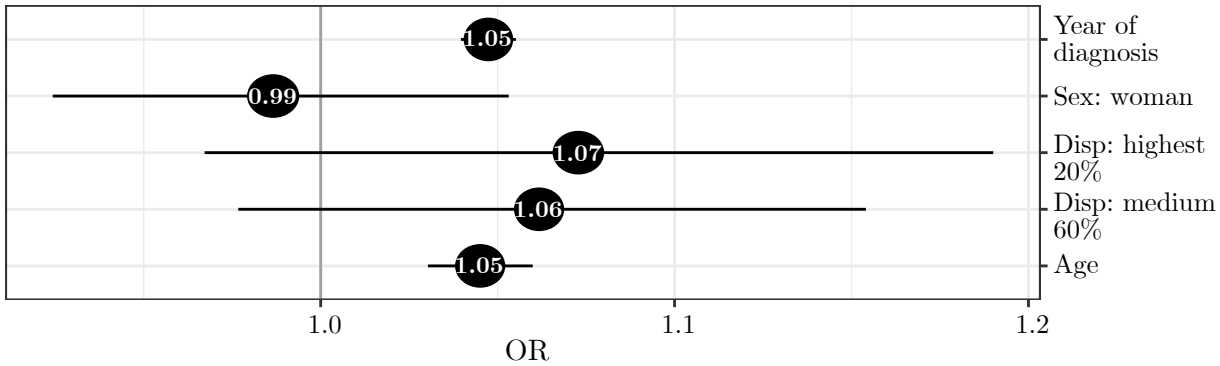

Figure 10: Odds ratio estimates of rectal cancer being diagnosed in stage localized with 95% confidence intervals. Reference group includes men from the poorest income group, who are 5 years younger and who were diagnosed with rectal cancer a year before.

When comparing the odds ratios of income groups according to year of diagnosis (Figure 11), we see that the two groups with the highest income have behaved rather similarly during our study period. As the odds ratios overlap, the differences between the two groups cannot be considered statistically significant.

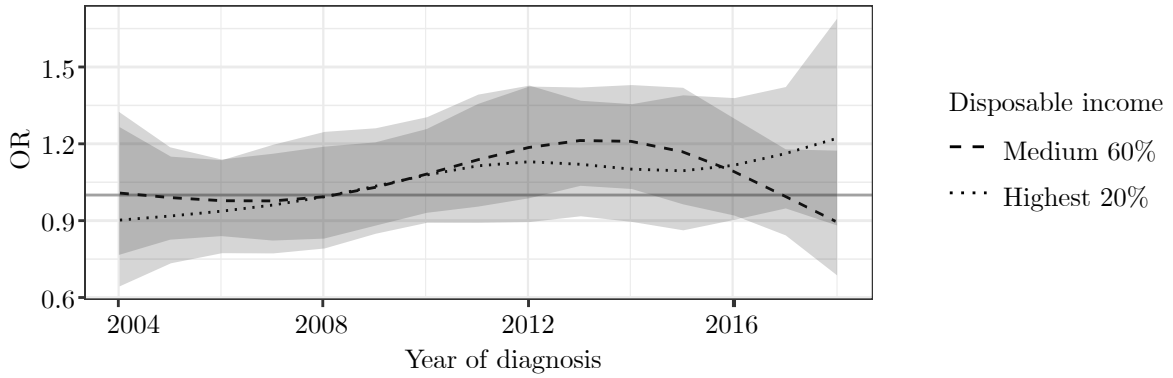

Figure 11: Odds ratio estimates of rectal cancer being diagnosed in stage localized from 2004 to 2018 with 95% confidence intervals. Reference group includes people from the poorest income group.

If we examine the changes of the log odds of disposable income levels for the mean age of rectal cancer patients, we can also see from Figure 12 that the trends of different income levels have been rather similar from 2004 to 2015. Although the differences between disposable income levels are not immense, the odds of being diagnosed with rectal cancer at an early stage have gotten better overall throughout our observed period of time.

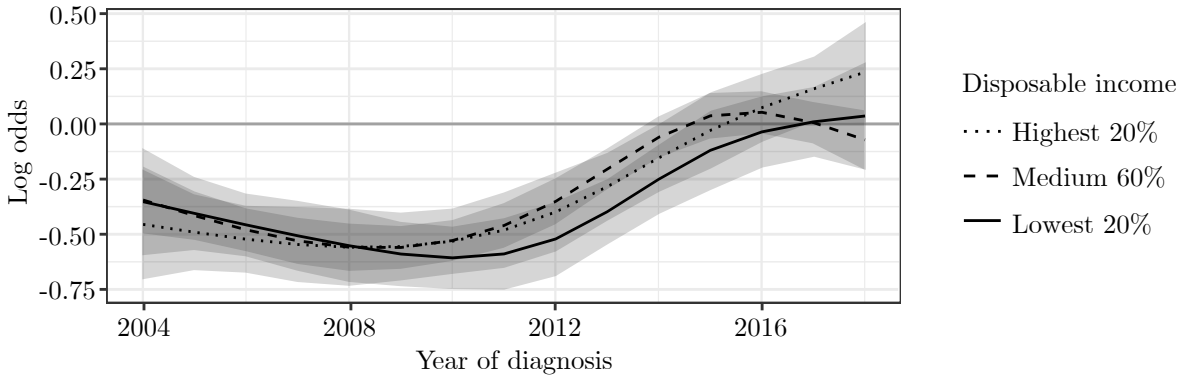

Figure 12: Log odds estimates of rectal cancer being diagnosed in stage localized by disposable income for people aged 69.5 years from 2004 to 2018 with 95% confidence intervals.

To summarise, having an education higher than just basic education slightly increased the odds of being diagnosed with early-stage rectal cancer. By disposable income, belonging to the medium 60% or wealthiest 20% groups might have also increased the odds a little, although it was not statistically significant. The odds of being diagnosed with early-stage rectal cancer have improved throughout our period of study for all educational and disposable income levels.

Conclusion

The aim of this bachelor's thesis was to examine whether socioeconomic factors had an impact on stage at diagnosis for colon and rectal cancers in Denmark from 2004 to 2018. 39 014 colon (ICD-10 C18) and 19 203 rectal (ICD-10 C20) cancer cases of patients over 18 years of age were analysed by disposable income and education to see whether the odds of being diagnosed at an early stage had improved or not. The analysis was based on two types of logistic regression models for both factors of interest – with and without restricted cubic splines.

The study showed that neither education nor disposable income had a statistically significant effect on at which stage colon cancer was diagnosed. Although the slight differences between educational and disposable income levels could not have been proven significant, we saw that the overall odds for being diagnosed with early-stage colon cancer improved from 2010 to 2015, but then began to decrease. This trend was most probably caused by the national colorectal screening programme introduced in 2014.

The analysis also showed that amongst rectal cancer patients there was a statistically significant association with education, showing that a person with a higher educational level tended to be diagnosed with rectal cancer at an earlier stage. The same conclusion could not have been done when studying disposable income. The odds of early-stage cancer diagnosis amongst rectal cancer patients seemed to increase, too, after the implementation of the screening programme, but did not show a clear decrease by the year 2018, as it did with colon cancer.

In conclusion, this population-based analysis suggests that income does not affect the staging of neither colon or rectal cancer. Education, on the other hand, has an impact on at which stage the cancer gets diagnosed amongst rectal, but not with colon cancer patients. As rectal cancer has more unmistakable symptoms than colon cancer, these results were as expected, because the patient has a better chance to react on symptoms. However, doing so is associated with socioeconomic position.

In light of this thesis, it might be interesting to study how income and education affect colon and rectal cancers' staging in Estonia and whether differences arise from the analysis done with Danish data. A similar study could also be done to study the impact of other socioeconomic factors such as cohabitation and housing statuses.

References

American Cancer Society (2020). *Colorectal Cancer Risk Factors*. URL: https://www.cancer.org/cancer/colon-rectal-cancer/causes-risks-prevention/risk-factors.html (30.03.2021).

American Society of Clinical Oncology (2021). *Colorectal Cancer: Introduction*. URL: https://www.cancer.net/cancer-types/colorectal-cancer/introduction/ (29.03.2021).

Causa, O., M. Hermansen, N. Ruiz, C. Klein, and Z. Smidova (2016). "Inequality in Denmark through the Looking Glass". In: 1341. URL: https://www.oecd-ilibrary.org/content/paper/5jln041vm6tg-en.

Danish Cancer Society (2021). *Screening for colon cancer*. URL: https://www.cancer.dk/international/english/screening-colon-cancer-english/ (29.03.2021).

Danmarks Statistik (2015). *AFSP1E*. URL: https://www.dst.dk/da/Statistik/dokumentation/Times/uddannelseregister/afsp1e (20.04.2021).

Egeberg, R., J. Halkjær, N. Rottmann, L. Hansen, and I. Holten (2008). "Social inequality and incidence of and survival from cancers of the colon and rectum in a population-based study in Denmark, 1994–2003". In: *European Journal of Cancer* 44.14, pp. 1978–1988. URL: https://doi.org/10.1016/j.ejca.2008.06.020.

Frederiksen, B. L., M. Osler, H. Harling, and T. Jørgensen (2008). "Social inequalities in stage at diagnosis of rectal but not in colonic cancer: a nationwide study". In: *British Journal of Cancer* 98.3, pp. 668–673. URL: https://doi.org/10.1038/sj.bjc.6604215.

Gauthier, J., Q. V. Wu, and T. A. Gooley (2020). "Cubic splines to model relationships between continuous variables and outcomes: a guide for clinicians". In: *Bone Marrow Transplantation* 55.4, pp. 675–680. URL: https://doi.org/10.1038/s41409-019-0679-x.

Hamilton, W. (2009). "The CAPER studies: five case-control studies aimed at identifying and quantifying the risk of cancer in symptomatic primary care patients". In: *British Journal of Cancer* 101.2, pp. 80–86. URL: https://doi.org/10.1038/sj.bjc.6605396.

Harrell, F. E. (2021). *Package 'rms'*. URL: https://cran.r-project.org/web/packages/rms/rms.pdf (17.05.2021).

James, G., D. Witten, T. Hastie, and R. Tibshirani (2013). *An Introduction to Statistical Learning with Applications in R*. Springer.

Larsen, M. B., S. Njor, P. Ingeholm, and B. Andersen (2018). "Effectiveness of Colorectal Cancer Screening in Detecting Earlier-Stage Disease—A Nationwide Cohort Study in Denmark". In: *Gastroenterology* 155.1, pp. 99–106. URL: https://www.sciencedirect.com/science/article/pii/S0016508518304050.

Mackenbach, J. P., I. Stirbu, A.-J. R. Roskam, M. M. Schaap, G. Menvielle, M. Leinsalu, and A. E. Kunst (2008). "Socioeconomic Inequalities in Health in 22 European Countries". In: *New England Journal of Medicine* 358.23, pp. 2468–2481. URL: https://doi.org/10.1056/NEJMsa0707519.

National Cancer Institute (2015). *Cancer Staging*. URL: https://www.cancer.gov/about-cancer/diagnosis-staging/staging/ (29.03.2021).

National Cancer Institute (2018). *Colon*. URL: https://www.cancer.gov/publications/dictionaries/cancer-terms/def/colon/ (29.03.2021).

O'Connell, J. B., M. A. Maggard, and C. Y. Ko (2004). "Colon Cancer Survival Rates With the New American Joint Committee on Cancer Sixth Edition Staging". In: *JNCI: Journal of the National Cancer Institute* 96.19, pp. 1420–1425. URL: https://doi.org/10.1093/jnci/djh275.

Wiking, M. (2016). *Why Danes Happily Pay High Rates of Taxes*. URL: https://www.usnews.com/news/best-countries/articles/2016-01-20/why-danes-happily-pay-high-rates-of-taxes (31.03.2021).

World Health Organization (2019). *International Statistical Classification of Diseases and Related Health Problems 10th Revision (ICD-10)*. URL: https://icd.who.int/browse10/2019/en (29.03.2021).

World Health Organization (2020). *Denmark: Cancer Country Profile 2020*. URL: https://www.who.int/cancer/country-profiles/DNK_2020.pdf (31.03.2021).

Non-exclusive licence to reproduce thesis and make thesis public

I, Kai Budrikas,

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to reproduce, for the purpose of preservation, including for adding to the DSpace digital archives until the expiry of the term of copyright,
   The impact of social inequality on stage at diagnosis of colon and rectal cancers in Denmark,
   supervised by Krista Fischer and Christian Dehlendorff.

2. I grant the University of Tartu a permit to make the work specified in p. 1 available to the public via the web environment of the University of Tartu, including via the DSpace digital archives, under the Creative Commons licence CC BY NC ND 3.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work until the expiry of the term of copyright.

3. I am aware of the fact that the author retains the rights specified in p. 1 and 2.

4. I certify that granting the non-exclusive licence does not infringe other persons' intellectual property rights or rights arising from the personal data protection legislation.

Kai Budrikas
18/05/2021