UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE ESTATÍSTICA E INVESTIGAÇÃO OPERACIONAL

Spatial distribution of the severity of lung cancer at diagnosis – is it related to socioeconomic factors and access to primary health care?

Mariana Reis Vieira

Mestrado em Bioestatística

Trabalho de Projeto orientado por:

Orientador: Profª Doutora Marília Antunes
Coorientador: Doutora Patrícia Soares

2021

Resumo

O cancro do pulmão é dos cancros mais fatais a nível mundial. As estimativas em 2018 para Portugal indicam que 4671 indivíduos morreram de cancro do pulmão, o que corresponde a 16.1% do total de mortes causadas por cancro. Estima-se que existam 5284 novos casos por ano, correspondendo a 9.1% de todos os cancros. A taxa de incidência para homens é 38.8 por cada 100000 habitantes enquanto que para mulheres é 12.6 por cada 100000 habitantes, o que corresponde a um aumento de 75%.

A elevada taxa de mortalidade neste tipo de cancro pode ser justificada pelo facto de se tratar de uma doença assintomática. Cancros em estadios avançados têm um prognóstico pouco favorável quando comparados com cancros detetados em estadios menos avançados, daí a importância de um diagnóstico precoce. O estadio determina a escolha de tratamento e representa a severidade do tumor, o que influenciará o tempo de sobrevivênia. A classificação TNM é um sistema de estadios criado com base em três critérios de informação: o tamanho do tumor primário (T), extensão para os nódulos linfáticos vizinhos (N) e extensão para orgãos distantes (M). De acordo com os exames de diagnóstico, a doença pode ser classificada como I, II, IIIA, IIIB or IV, sendo um indicador da severidade da doença.

A nova campanha , *Treatment for All*, da União para o Controlo Internacional do Cancro tem como objetivo reduzir a morte prematura de cancro e promove o acesso equitativo para o tratamento e bem-estar. As condições socioeconómicas são alguns dos fatores que podem comprometer o acesso aos cuidados de saúde. Portanto, o principal objetivo deste estudo foi perceber se os fatores socioeconómicos e o acesso aos cuidados de saúde estão associados com o estadio em que o cancro é diagnosticado.

A informação foi recolhida pelo Registo Oncológico Regional Sul (ROR-Sul), que inclui as regiões de Lisboa e Vale do Tejo, Alentejo, Algarve e Região Autónoma da Madeira. O conjunto de dados tinha incluído 2266 pacientes diagnosticados com cancro do pulmão em 2013 e 2014.

As variáveis incluídas foram o género, idade, concelho de residência, distrito de residência, morfologia, lateralidade, estadio ao diagnóstico e estado vital. As variáveis socioeconómicas foram extraídas a partir do INE e PORDATA. Através da revisão de literatura, foram identificados alguns indicadores que caraterizam as condições socioeconómicas, bem como as de acesso aos cuidados de saúde.

Os dados foram modelados aplicando o modelo de regressão ordinal e o modelo misto de regressão ordinal, usando o concelho de residência como um efeito aleatório, que corresponde à variável que liga o conjunto de dados originais aos indicadores socioeconómicos e de acesso aos cuidados de saúde. O termo aleatório explicará as diferenças entre os concelhos e reduz a componente por explicar do modelo sem um termo aleatório.

A correlação linear foi analisada para evitar a inclusão de variáveis independentes fortemente correlacionadas. A variável escolhida entre o par fortemente correlacionado era a mais informativa, excluindo aquela que, sendo menos informativa, estava associada à que foi incluída. A influência de cada uma das variáveis foi analisado de acordo com o odds ratio (OR).

Considerando o sinal dos coeficientes de regressão, os resultados do modelo múltiplo sem termo aleatório indicaram que maior número de médicos por cada 1000 habitantes (OR 0.974, 95% CI: 0.942

- 1.008), idades avançadas (OR 0.996, 95% CI: 0.989 - 1.004) e maior número de beneficiários por cada 1000 habitantes (OR 0.998, 95% CI: 0.993 - 1.004) aparentam favorecer estadios mais baixos. Um maior rendimento anual (OR 1.003, 95% CI: 0.949 - 1.060) e um maior número de atendimentos por cada 1000 habitantes (OR 1.005, 95% CI: 0.995 - 1.016), aparentam contribuir para um diagnóstico em estadios avançados. O impacto do género variou de acordo com a categoria da variável resposta. Incluindo o termo aleatório, os resultados também indicaram que um elevado número de médicos por cada 1000 habitantes (OR 0.971, 95% CI: 0.880 - 1.073), uma idade avançada (OR 0.996, 95% CI: 0.988 - 1.004) e um maior número de beneficiários por cada 1000 habitantes (OR 0.998, 95% CI: 0.988 - 1.009) aparentam favorecer estadios menos avançados. Um elevado rendimento anual (OR 1.008, 95% CI: 0.942 - 1.078) e um maior número de atendimentos por cada 1000 habitantes (OR 1.007, 95% CI: 0.988 - 1.026) aparentam contribuir para um diagnóstico em estadios avançados. Ao contrário do modelo sem termo aleatório, o efeito do género não varia de acordo com a severidade da doença. Com base no sinal do seu coeficiente de regressão, a possibilidade de um homem ser diagnosticado num estadio avançado era menor que uma mulher (OR 0.866 , 95% CI: 0.572 - 1.312). Apesar da variância associada ao termo aleatório (concelho de residência) tenha sido próxima de 1, a diferença entre estas regiões foram estatisticamente significativas no que diz respeito à severidade do estadio ao diagnóstico. A análise geoespacial mostrou que uma região do Centro tinha menor possibilidade de diagnóstico em estadios superiores. Na Região Autónoma da Madeira, a possibilidade de diagnóstico em estadios superiores era maior.

Os resultados dos modelos múltiplos não encontraram evidências de associação entre as condições socioeconómicas e o acesso aos cuidados de saúde e a severidade do cancro do pulmão. O trabalho futuro deve passar pela recolha de mais informações individuais sobre o paciente, como estado civil, hábitos tabágicos, alimentção, mas também condições económicas e de acesso aos cuidados de saúde, como ter médico de família, proximidade de centros de saúde, facilidade para sair do trabalho, cobertura de seguro, etc.

Palvras-chave: cancro do pulmão, dados ordinais, condições socioeconómicas, modelo cumulativo, modelo de odds proporcionais, modelo de odds proporcionais parciais, modelo misto cumulativo, efeitos alearórios

Abstract

Lung cancer is the most lethal type of cancer worldwide. The estimates for Portugal in 2018 indicate that 4671 individuals died of lung cancer, corresponding to 16.1% of total cancer deaths, with 5284 new cases estimated per year, corresponding to 9.1% of all cancers. The incidence rate for males is 38.8 per 100000 inhabitants whereas for females is 12.6 per 100000 inhabitants, which corresponds to an 75% increase.

The high mortality rate of this type of cancer can be attributed to the fact that it is an asymptomatic disease, which delays diagnosis. Cancers in more advanced stages have reduced favourable prognosis compared to cancers detected in earlier stages, hence the importance of early diagnosis. The stage determines the choice of treatment and represents the severity of the tumour, which will influence survival time. TNM classification is a staging system created based on three information criteria: the size of the primary tumor (T), the spread to nearby lymph nodes (N) and the spread to distant organs (M). According to the diagnostic exams, the disease can be classified as I, II, IIIA, IIIB or IV, being an indicator of the severity of the disease.

The new campaign, *Treatment for All*, of the Union for International Cancer Control (UICC) aims to reduce premature mortality from cancer and promote equitable access to treatment and care. Socioeconomic conditions can compromise access to primary health care. Therefore, the main aim of this study was to understand if socioeconomic factors and access to primary health care are associated with the stage at which the cancer is diagnosed.

Data were collected from the Southern Portugal Cancer Registry (ROR-Sul), which includes the regions of Lisbon and the Tagus Valley, Alentejo, Algarve and Autonomous Region of Madeira. The dataset had included 2266 patients diagnosed with lung cancer in 2013 and 2014.

The variables included in the original dataset were gender, age, residence county, residence district, morphology, laterality, stage at diagnosis and vital status. Socioeconomic variables were downloaded from the INE and PORDATA. Through a literature review several indicators characterizing the socioeconomic conditions as well as the access to healthcare conditions were identified.

The data were modelled applying the ordinal regression model and the ordinal regression mixed model using the residence county as a random effect, which corresponds to the variable that links the original dataset to the socioeconomic and access healthcare indicators. The random term will explain the differences between counties and reduce the unexplained component of the model without a random term.

The linear correlation was analysed to avoid the inclusion of strongly correlated independent variables. The variable chosen among the strongly correlated pair was the most informative, excluding the one that, being less informative, was associated with the one that was included. The influence of each variable was analysed according to the odds ratio (OR).

Considering the sign of the regression coefficients, the results of the multivariable model without random term indicated that higher number of doctors per 1000 inhabitants (OR 0.974, 95% CI: 0.942 -

1.008), higher age (OR 0.996, 95% CI: 0.989 - 1.004) and higher number of welfare recipients per 1000 inhabitants (OR 0.998, 95% CI: 0.993 - 1.004) appeared as favouring lower stages. A higher annual income (OR 1.003, 95% CI: 0.949 - 1.060) and a higher number of attendances per 1000 inhabitants (OR 1.005, 95% CI: 0.995 - 1.016), appeared as contributing to a diagnosis in higher stages. The impact of gender varied according to the category. Including the random term, the results also indicated that a higher number of doctors per 1000 inhabitants (OR 0.971, 95% CI: 0.880 - 1.073), a higher age (OR 0.996, 95% CI: 0.988 - 1.004) and a higher number of welfare recipients per 1000 inhabitants (OR 0.998, 95% CI: 0.988 - 1.009) appeared favouring lower stages. A higher annual income (OR 1.008, 95% CI: 0.942 - 1.078) and a higher number of attendances per 1000 inhabitants (OR 1.007, 95% CI: 0.988 - 1.026), appeared as contributing to a diagnosis in higher stages. Unlike the model with no random term, the effect of gender does not vary according to the severity of the disease. Based on the sign of its regression coefficient, the odds of a male being diagnosed at a later stage was less than a woman (OR 0.866, 95% CI: 0.572 - 1.312). Although the variance associated with the random effect (residence county) was close to 1, the difference within regions were statistically significant regarding the severity of stage at diagnosis. The geospatial analysis has shown that a region in the Center had a lower possibility of having a diagnosis at higher stages. In the Autonomous Region of Madeira, the possibility of having a diagnosis at higher stages was higher.

The multivariable models results found no evidence of a statistically significant association between socioeconomic conditions and access to healthcare, as they were measured, and lung cancer severity. Future work should collect more individual information about the patient, such as marital status, smoking habits, diet, but also economic conditions and conditions accessing healthcare, such as having a family doctor, proximity to health centres, ease of leaving work, insurance coverage, etc.

Keywords: lung cancer, ordinal data, socioeconomic conditions, cumulative link model, proportional odds model, partial proportional odds model, cumulative link mixed model, random effects

Acknowledgment

This project was a huge challenge in my life. The support of some people was crucial, especially during this phase.

First of all, I would like to thank Professor Marília Antunes and Patrícia Soares for the guidance on this project. They supervised my project, being always available to clarify any question.

To the special person, who followed and supported me unconditionally and in this challenge was no exception. It was not easy, but Jorge was always with me, as he has been in the last years of my life.

I would like to thank my family. To my parents, sister and brother in law. They supported me not only during this phase but also at all times in my life. Some special people are also included: Carla, João and Armando, thanks for the concern, affection and appreciation.

Special thanks to the team I have been part of since 2019, especially to Filipa. She was always available to help with whatever was needed to complete this chapter of my life.

I would like to thank all my friends, especially Miguel, Alexandre, Nuno, André, Catarina and Inês. They helped me during this phase emotionally.

I will never forget the support that these people, in different ways, gave me the strength to conclude this challenge.

Índice

List of Figures

List of Tables

Abbreviations and Acronyms

CI: Confidence interval

ICD-O: International Classification of Diseases for Oncology

INE: National Institute of Statistics

NUTS: Nomenclature of territorial units for statistical purposes. The nomenclature is subdivided into 3 levels (NUTS I, NUTS II, NUTS III), defined according to population, administrative and geographical criteria.

> **NUTS I:** Portugal is divided into the following units: Continent, Lisbon Metropolitan Area, Autonomous Region of Madeira and Autonomous region of Azores
>
> **NUTS II:** Portugal is divided into the following units: North, Center, Alentejo, Algarve, Lisbon Metropolitan Area, Autonomous Region of Madeira and Autonomous region of Azores
>
> **NUTS III:** Portugal is divided into inter-municipal entities (23 territorial units), Autonomous Region of Azores and Autonomous Region of Madeira

NSCLC: Non-small cell lung cancer

OR: Odds ratio

ROR-Sul: South Regional Cancer Registry

Chapter 1

Introduction

This section will describe lung cancer epidemiology, diagnosis and treatment, including the criteria defining the lung cancer stage. A brief review of the risk factors associated with lung cancer disease and severity will also be presented. An overall description of the project will be made, considering two approaches: individual and area-based information. The area-based information will be crucial to achieving the aim of the project since socioeconomic conditions and access to healthcare are not presente dat the individual level. The chapter ends with a description of the main objectives within this project.

1.1 Lung Cancer

Lung cancer results from an abnormality in the basic unit of life, the cell. Normally, the body maintains a system of checks and balances on cell growth so that cells divide to produce new cells only when new cells are needed. Disruption of this system of checks and balances on cell growth results in an uncontrolled division and proliferation of cells that eventually form a mass known as a tumor.

Lung cancers are generally divided into two main categories: small cell lung cancer (SCLC) and non–small cell lung cancer (NCSLC) [30, 32]. SCLC are clinically aggressive and they are usually centrally located with extensive mediastinal involvement and associated with early extrathoracic metastases, including paraneoplastic syndrome. Despite a good response to chemotherapy, SCLC are often diagnosed at a late stage and patients have poor prognosis. Histologically, NSCLC are further divided into adenocarcinoma, squamous cell carcinoma, and large cell carcinoma. These three categories share treatment approaches and prognoses but have distinct histologic and clinical characteristics. It should be noted that 85% are NSCLC. Histologically, adenocarcinomas are heterogeneous peripheral masses that metastasize early, and often occur in patients with underlying lung disease. Overall, about 40% of lung cancer are adenocarcinomas. Typically, squamous cell carcinomas are centrally located endobronchial masses that may present with hemoptysis, postobstructive pneumonia or lobar collapse. Unlike adenocarcinomas, squamous cell carcinomas generally metastasize late in the disease course. About 25% to 30% of lung cancers are squamous cell carcinomas. Large cell carcinomas can begin in any part of the lung. They tend to grow and spread quickly, which can make it harder to treat. A subtype of large cell carcinoma, known as large cell neuroendocrine carcinoma, is a fast-growing cancer similar to SCLC. This type of tumor accounts for about 10% to 15% of lung cancers. Figure 1.1 illustrates the different types of cancer.

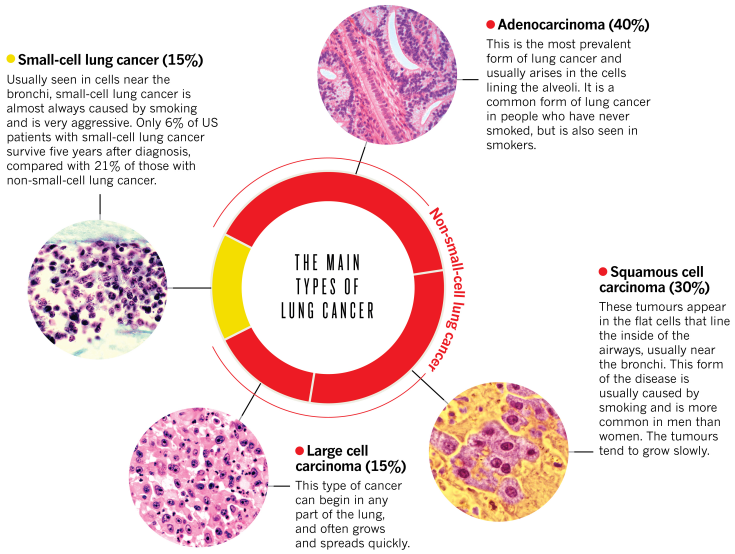

Figure 1.1: Classification of the different types of lung cancer. [7]

According to the World Health Organization (WHO), in 2013 the worldwide mortality rate due to lung cancer was 38.3 per 100000 inhabitants, that is, in 100000 inhabitants, about 38 died due to this pathology per year. For every 100000 inhabitants, about 16 women die and 75 men die per year [18].

More recently, the estimates for Portugal in 2018 indicate that 4671 individuals died from lung cancer, corresponding to 16.1% of total cancer deaths, with 5284 new cases estimated, corresponding to 9.1% of all cancers. The same trend in gender was observed in Portugal, with an incidence rate of 38.8 for males and 12.5 for females, per 100000 inhabitants per year [24]. The estimated number of new cases was 3998 for males, while for females it was 1286 (see Appendices - Figures 1 and 2).

In 2013 and 2018, the Portuguese mortality rate was lower than the European mortality rate [23, 20]. In 2018, 20% of total cancer deaths in Europe (1943478 cases) were caused by lung cancer. The European incidence of this disease was 11.1%, in which, in a total of 4229662 new diagnosed cancer cases, 470039 correspond to the lung [23]. In Europe, the incidence between genders was also different, the associated incidence to the males and females patients was 44.3 and 18.3 per 10000 inhabitants, respectively (see Appendices - Figures 5 and 6).

The worldwide mortality rate of lung cancer in 2018 was 18.4%, lower than Europe but higher than Portugal [25]. In 18078957 of diagnosed new cases with cancer, 11.1% corresponded to the lung. Worldwide, the incidence in male patients was 54% higher comparing to the female patients (see Appendices - Figures 5 and 6).

Overall, comparing the incidence rate of lung cancer, Portugal (9.1%) is below the levels of Europe (11.1%) and worldwide (18.4%). Regardless of the geographic region, Portugal, Europe or worldwide, there is a clear difference in the incidence rate between male and female patients.

1.1.1 Diagnosis

The common tests used to diagnose lung cancer are image tests, tissue sample (biopsy), among others [8]. Based on the results and the staging criteria, the severity of the lung cancer is assessed.

1.1.2 Stage

After the patient is diagnosed with NSCLC, doctors will assess if the tumour has metastasised, its size and the involvement of lymph nodes. This process is called staging. The stage of cancer describes how serious the cancer is and how to approach treatment [33].

The stages range from I through IV. As a rule, the lower number corresponds to the initial phase of the tumor. A higher number means cancer has spread to more regions.

The staging system widely used for cancer is the TNM system, which is based on three key pieces of information: the size and extent of the primary tumor **(T)**; the spread to nearby lymph nodes **(N)**; the spread (metastasis) to distant sites **(M)**. Non-small cell lung cancer staging can be classified as follows:

Stage I: The dimension of the tumor is lower than 5cm, invaded the deeper tissue of the lung without affecting nearby lymph nodes or the chest wall.

Stage II: The dimension of the tumor is lower than 7cm and there is the possibility of invasion of the adjacent lymph nodes; or its dimension is lower than 5cm but it has already invaded adjacent tissues, such as the wall, diaphragm, pleura, bronchi, or tissues around the heart.

Stage IIIA:

- The dimension of the tumor is 5cm or smaller and cancer has spread to lymph nodes on the same side of the chest as the primary tumor. The lymph nodes with cancer are around the trachea or where the trachea divides into the bronchi. One or more of following will also be present:
  - Cancer has spread to the main bronchus, but has not spread to carina.
  - Cancer has spread to the innermost layer of the membrane that covers the lung.
  - Part of the lung or the whole lung has collapsed or has developed pneumonitis.
- Cancer has spread to lymph nodes on the same side of the chest as the primary tumor. The lymph nodes with cancer are in the lung or near the bronchus. One or more of following will also be present:
  - The dimension of the tumor is larger than 5cm but not larger than 7cm.
  - There are one or more separate tumors in the same lobe of the lung as the primary tumor.
  - Cancer has spread to the membrane that lines the inside of the chest wall, chest wall, the nerve that controls the diaphragm or outer layer of tissue of the sac around the heart.
- Cancer may have spread to lymph nodes on the same side of the chest as the primary tumor. The lymph nodes with cancer are in the lung or near the bronchus. One or more of following will also be present:
  - The dimension of the tumor is larger than 7cm.
  - There are one or more separate tumors in a different lobe of the lung with the primary tumor.
  - The tumor is of any size and cancer has spread to the trachea, carina, esophagus, breastbone or backbone, diaphragm, heart, major blood vessels that lead to or from the heart or nerve that controls the larynx.

Stage IIIB:

- The dimension of the tumor is 5cm or smaller. Cancer has spread to lymph nodes above the collarbone on the same or opposite side of the chest as the primary tumour. One or more of following will also be present:

  - Cancer has spread to the main bronchus, but has not spread to the carina.
  - Cancer has spread to the innermost layer of the membrane that covers the lung.
  - Part of the lung or the whole lung has collapsed or has developed pneumonitis.

- The tumor may be of any size and cancer has spread to lymph nodes on the same side of the chest as the primary tumor. The lymph nodes with cancer are around the trachea or where the trachea divides into the bronchi. Also, one or more of the following is found:

  - There are one or more separate tumors in the same lobe or a different lobe of the lung with the primary tumor.
  - Cancer has spread to the membrane of the chest wall inside, chest wall, the nerve that controls the diaphragm, outer layer of tissue of the sac around the heart, trachea, carina, esophagus, breastbone or backbone, diaphragm, heart, major blood vessels that lead to or from the heart or nerve that controls the larynx.

Stage IV: The cancer is present in both lungs or has metastasised to more distant organs such as the brain, bones or liver, or there is the presence of lung cancer cells in the fluid located between the two layers of the pleura.

Figure 1.2, illustrates the different stages of lung cancer.

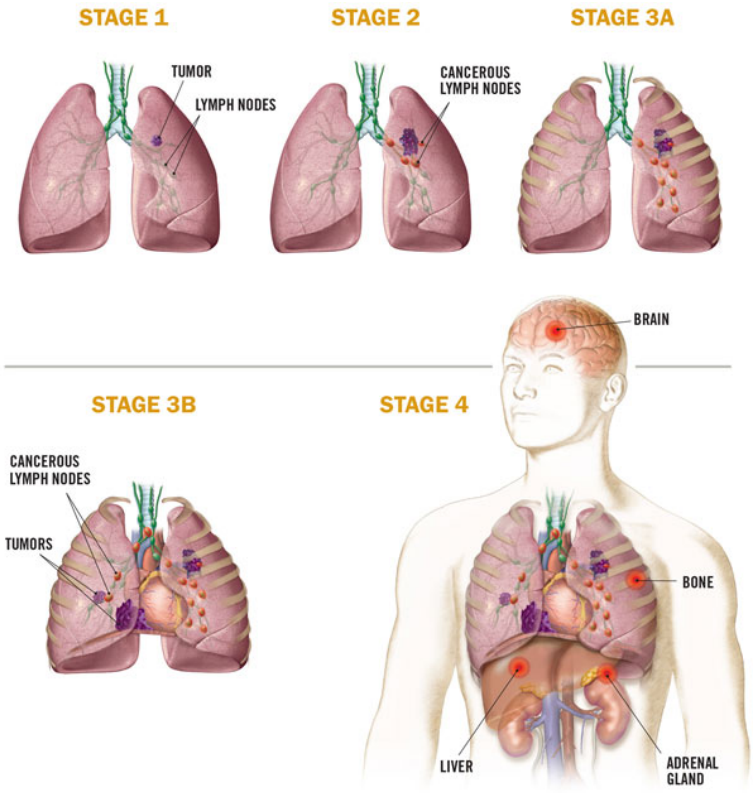

Figure 1.2: Staging of non-small cell lung cancer. [2]

1.1.3 Treatment and Prognosis

The decision about which treatments will be appropriate for a given individual must consider the location and extent of the tumor (stage of the cancer), and the overall health of the patient [30, 34].

Treatment for lung cancer involves primarily surgical removal of the tumour, chemotherapy, radiation therapy, and combinations of these treatments. Therapy may be prescribed that is intended to be curative (removal or eradication of cancer) or palliative (measures that are unable to remove the tumour but can reduce pain and suffering). In such cases, the therapy, referred to as adjuvant therapy, is added to enhance the effects of the primary therapy. An example of adjuvant therapy is chemotherapy or radiotherapy, which is administered after surgical removal of a tumor in an attempt to kill any tumor cells that remain following surgery.

- **Surgery:** Surgery is generally performed when cancer has not spread beyond the lung, hence stage I and sometimes stage II. About 10%-35% of lung cancers can be removed surgically, but removal does not always result in total remission.
- **Radiation:** Radiation therapy is used to kill dividing cancer cells. This treatment may be given as curative therapy, palliative therapy, or as adjuvant therapy combined with surgery or chemotherapy.

  This therapy generally only shrinks a tumor or limits its growth when given as a sole therapy, yet in 10%-15% of people it leads to long-term remission and palliation of the tumour. Combining radiation therapy with chemotherapy can further prolong survival when chemotherapy is administered.
- **Chemotherapy:** Chemotherapy refers to the administration of drugs that stop the growth of cancer cells by killing them or preventing them from dividing. Chemotherapy may be given alone or as adjuvant therapy to surgery or radiotherapy.
- **Immunotherapy:** Immunotherapy may be an effective option for some patients with advanced lung cancers. Immunotherapy drugs work by strengthening the activity of the immune system against tumor cells. These drugs are inhibitors that target checkpoints or areas that control the immune response and promote the immune response.

After diagnosis, the treatment is given according to the stage at diagnosis. The treatment may be simple for early stages, like surgery, and there is a possibility to eradicate the tumour. For later stages, the procedure may be more evasive and aggressive without guarantees of total tumor removal.

For people with stage I NSCLC, the five-year survival rate is between 60% and 70%. For the diagnosed patients with stage II lung cancer, the five-year survival rate is about 40% - 50%. The five-year survival rate for patients with stage IIIA NSCLC is between 10% and 30% depending if the tumour is resectable or not. For stage IIIB, the five-year survival is between 10% and 20%. Sadly, for people with stage IV NSCLC, the available measure is two-year survival because the patients diagnosed in stage IV does not survive 5 years. So, the two-year survival for these patients is between 10% and 20%.

1.2 Lung Cancer Risk Factors

Lung cancer is a malignant neoplasm with a high mortality rate. Several risk factors have been studied associated with lung cancer disease and survival [3, 11, 17, 35]. One of the commonly associated factors is the smoking habits. The effect of smoking is directly associated with the number of cigarettes smoked per day, the duration of smoking, the age of starting smoking, and even passive smokers have some risk [21, 35]. The epidemiological evidence supports a causal association between secondhand exposure to

cigarette smoke and lung cancer risk in nonsmokers [21]. The risk of a nonsmoker individual being diagnosed with lung cancer is 20%-30% higher if their spouse/partner is a smoker than a nonsmoker. The effect of involuntary smoking appears to be present for household exposure, mainly from spousal and workplace exposure.

Another factor associated with lung cancer is indoor air pollution, specifically coal burning in poorly ventilated houses, burning of wood and other solid fuels, and fumes from high temperatures cooking using unrefined vegetable oils. In several regions of Asia, indoor air pollution is a major risk factor for lung cancer in never-smoking women. [21] In Europe, a positive association between indicators of indoor air pollution and lung cancer risk has been reported [16]. Outdoor air pollution is considered a lung carcinogen in humans by the International Agency for Research [17]. However, the studies exploring the association between air pollutants and lung cancer have limitations since air pollution is measured using proxies, such as the number of inhabitants in the community of residence and residing near a major pollution source. Nevertheless, outdoor air pollution is another important risk factor for lung cancer.

Diet and alcohol are also indicators that have been studied. According to case-control studies, a diet rich in vegetables and fruits may exert some protective effect against lung cancer. A high intake of meat, in particular fried or well-done red meat, may increase the risk of lung cancer. Regarding the alcohol consumption, there is an article that reveals that it is associated with an increased risk of lung cancer. Alcohol consumption was categorized as follows: 0, 0.1-12, 12.1-24 and greater than 24g/day and for these categories, the incidence rates of lung cancer were 7.4, 13.6, 16.4 and 25.2 per 10000 person - years, respectively [9]. However, after adjustment for smoking status and other major risk factor, alcohol consumption was no longer statistically significantly associated with the risk of lung cancer [9]. Since the correlation between alcohol consumption and tobacco smoking in many populations, it is difficult to evaluate the contribution of alcohol to lung carcinogenesis while properly controlling for the potential confounding effect of tobacco.

Gender appears to be an important independent risk factor for developing lung cancer. The worldwide lung cancer incidence and mortality rates show higher values for men than for women. This trend is observed in all evaluated continents – Africa, Asia, America, Europe and Oceania. Some studies report that the risk of all major histological types of lung cancer is increasing more rapidly for women than for men [12]. These differences can be related to confounding factors, such as the income [26] or smoking habits [12].

Family history of lung cancer and genetic mutations also increase the risk of developing lung cancer [17].

1.3 Impact of Socioeconomic Conditions

Several socieconomic conditions, such as income, education, occupational exposure or household conditions, may increase the risk of lung cancer [3]. People with lower socioeconomic conditions have the highest incidence rates. The position in societal hierarchies is defined by socioeconomic status and, generally it is assessed by these three independent different dimensions: education, occupation and income. Education can determine the occupational opportunities and earning potential. For people with higher education level, it is easier to access information and resources to promote health [3]. Regarding income, there is a clear relation between income and health. People with higher income can access better life conditions: nutrition, housing and/or schooling. For survival, especially during the first semester since diagnosis of lung cancer, income may be a stronger marker of vulnerability than education, because it is a measure of overall resources during an especially vulnerable period, when access to care is of major

importance. For all these reasons it is considered that "the most fundamental cause for health disparities are socioeconomic disparities" [3].

Socioeconomic conditions can be measured individually or based on socioeconomic statistics available for the residential area of each patient. This last type of data is used in ecological approaches or area-based studies, where the unit of observation is the residential area. A systematic review assessed the effect of socioeconomic conditions, measured at the individual and area-based level, on lung cancer survival. Overall, 94 studies were included, in which 23 used measures socioeconomic status on an individual level and 71 area-based. The main objective is to provide a summary on the current literature on socioeconomic differences in lung cancer survival, focusing on the impact of aggregation and individualization level.

The education can be an important factor to determine the risk of death and, combining the individual income, the results revealed no significant associations between this factor and lung cancer survival. However, with area-based data, an association was found, in which the lung cancer survival is lower in regions with less educated patients.

Also, the impact of income has been evaluated in this type of studies. The results showed that the lung cancer survival is lower for patients whose income is lower. Most of area-based studies revealed that chance to survive is poor for the lowest income group compared to the highest group.

The occupation was another favor evaluated in this set of studies, where the lowest survival chance is not associated to the lowest socioeconomic conditions for occupational groups. However, including the education level, the main conclusion was that the risk of death in patient with high-level non-manual occupation and medium education is lower compared to the low educated patient with manual occupation.

Health insurance coverage is an important indicator of access to care and a possible cause of disparities in lung cancer outcomes. There is an article in which the association of insurance status with measures of access to care was evaluated [31]. The results suggested that patients with no insurance had poorer lung cancer outcomes, including higher incidence rate, later stage at diagnosis, and poorer survival than patients with insurance [31]. Some of the disparities may be a secondary to residual confounding from other health behaviours, but data suggested that patients with lung cancer without insurance do poorly because access to care is limited [31].

Living conditions were also associated with lung cancer development. Cancers in patients who live alone were more likely to be diagnosed at a higher stage and were less likely to receive appropriate treatment [15]. The percentages of patients who were married, single, divorced, and widowed who received no cancer treatment were 30%, 33%, 38%, and 39%, respectively. The percentages of patients who underwent surgery based on martial status were 41%, 39%, 35% and 41%. For chemotherapy, the percentages were 34%, 27%, 31% and 19%, respectively. Differences in radiation administration were 27%, 26%, 24% and 19%, respectively, with widowed patients having the lowest rate of radiation administration and chemotherapy. Also, the conditions in which the patient lives also influence the evolution of lung cancer. Cancers in patients who live alone are more likely to be diagnosed at a later stage, as well as they are less likely to receive more appropriate treatment. There are evidences that living together are associated with better cancer outcomes. Structural and emotional support are crucial throughout diagnosis, treatment and rehabilitation. Similarly, regardless of sex, married individuals had better 5-year survival than non-married individuals [15]. The positive effects of marriage on lung cancer were attributed to strong social networks, which can positively influence neuroimmune pathways and health behaviours [28].

There is no data regarding the effect of socioeconomic conditions and access to healthcare on lung cancer severity in Portugal. Hence, the motivation for this thesis was to explore this theme.

1.4 Case Study

Lung cancer is one of the most common malignant tumours with the highest mortality rate of the total cancer deaths [22]. Among other factors, the asymptomatic characteristics of the disease contribute to the fact that most cases are diagnosed at an advanced stage which decreases the survival probability of patients with lung cancer. Lung cancer is considered an asymptomatic disease because it may take years to develop symptoms or it may not be detected until the disease is at an advanced stage or simply because the symptoms may be similar to symptoms from other causes - cough, chest pain, hoarseness, *etc*.

Lung cancer corresponds to the most common cause of cancer death worldwide for men and the second most common for women. Several factors impact lung cancer severity, some of which are socioeconomic conditions, smoking habits and gender.

This project aims to study the association between the severity of the disease at diagnosis and the socioeconomic conditions the patient is exposed, and access to healthcare.

The original dataset contains demographic information regarding each patient such as gender, age, residence area (county and district) and status (dead or alive). Data about income, education and occupation were obtained for each county through an external source - INE or PORDATA. This information was merged with the original dataset through the residence county of each patient. The collected information is chosen based on literature review and the available information between the two external sources.

With spatial data, the severity of the disease is modelled according to the characterization of each patients' area of residence and not by their own characteristics. Thus, if significance differences are found, preventive measures to combat the high mortality in that specific area can be taken.

1.5 Objective and Analysis Plan

The main objective of this project is to study the association between the stage at diagnosis and socioeconomic and access to healthcare conditions at the time of diagnosis in a group of patients diagnosed with non-small cell lung cancer in 2013 and 2014 in the southern region of Portugal.

The first step of this project concerns data collection and treatment. The original dataset includes neither socioeconomic conditions nor access healthcare indicators. So, this type of information will be merged with the original data using geographical level variables. In the presence of correlated variables and to avoid variance inflation, contributing to a parsimonious model, only one variable will be chosen. Here, two main sets are considered: set of socioeconomic variables and set of access to healthcare variables. Exploratory analysis will be done for both demographics and the variables selected previously. After knowing how the sample is distributed and the behavior between some variables, the statistical inference follows. Considering that the stage at diagnosis is an ordinal variable with more than 2 categories, ordinal regression models will be used. One model will have a random effect to accommodate differences between regions because socioeconomic data and access to healthcare were available at region-based level.

The statistical software used in this project is R version 4.0.5 and the required packages are readxl to import the dataset; tidyverse to data representations and data managements; ordinal to implement the cumulative link model as well as the cumulative link mixed model; ggplot2 to create graphics; rgdal to bindings for the 'geospatial' data abstraction library; maptools to manipulating geographic data; maps to draw geographical data and mapproj to converts latitude and longitude into projected coordinates. The last four packages were used to build the maps in *R*.

1.6 Overview

In chapter 2, both demographic data and socioeconomic and access to healthcare indicators used in this project are described, and the correlation between variables is measured.

Chapter 3 describes the methodology used in this project - ordinal logistic regression and its extension, including a brief overview of the multinomial regression.

Chapter 4 contains the exploratory analysis and the model results in which it is possible to analyse the impact of each explanatory variable on the severity of lung cancer. Both univariable and multivariable analysis are included.

In chapter 5, the results are discussed, and the two applied models (ordinal regression model and ordinal regression mixed model) are compared. Limitations, future work and alternative approaches to model this type of outcomes are also discussed.

Chapter 2

The data

This chapter describes the dataset obtained from the South Regional Cancer Registry (ROR-Sul) and the variables used for the analysis. Since the project aims to study the association between the severity of NSCLC: non-small cell lung cancer and the socioeconomic conditions and access to health care, several variables were downloaded from INE and PORDATA. Some variables were transformed - to eliminate the population scale factor - and merged with the main dataset. Finally, due to the similarity of some socioeconomic variables, the linear correlation among them was calculated. In the presence of linear correlation, only one variable was selected. Since the response variable is ordinal with more than two categories, the ordinal regression model was applied. Additionally, in order to accommodate the differences within the regions, the ordinal regression mixed model was applied.

2.1 Study Population

The South Regional Cancer Registry (ROR-Sul) collected the data used in this project, and it is the base of this retrospective cohort study. ROR-Sul is a population-based cancer registry, established in 1988, and covers approximately 50% of the Portuguese territory - Lisbon and the Tagus Valley (Leiria, Lisbon, Santarém and Setúbal districts), Alentejo (Portalegre, Évora and Beja districts), Algarve (Faro district) and the Autonomous Region of Madeira are (see Figure 2.1). About 128 counties of Portugal are included in the area that covers the ROR-Sul. On the 1^{st} January of 2018, according to the Decree Law n.o 53/2017 of the 14^{th} July, the National Cancer Registry (RON) was established, centralizing the information of all patients diagnosed with cancer in Portugal in one database. However, only data from ROR-Sul was included in the analysis since it was internally validated for scientific purposes.

Figure 2.1: Area belonging to the ROR-Sul by counties.

2.1.1 Sample and Data Collection

The project included patients diagnosed with NSCLC cancer in 2013 and 2014 (from 01-01-2013 to 31-12-2014) living in the ROR-Sul residence area were included in the cohort. The data included in the study was been externally validated by one of the researchers.

In this study the inclusion criteria were the following:

- adult patients (> 18 years old)
- patients who lived in ROR-Sul area
- histopathological or cytological diagnosis of non-small cell lung cancer according to ICD-O 3^{rd} revision
- diagnosis date in 2013 or 2014

The exclusion criteria were the following:

- patients without stage at diagnosis registered
- patients without laterality registered or registered as bilateral
- patients with foreign nationality

2.1.2 Original Data

An anonymised identification number was created - ID for each patient. Table 2.1 contains the description of the clinical and sociodemographic variables included in the dataset. All the variables correspond to individual data that characterizes the patient, such as gender, age, residence district, stage

at diagnosis, etc. The table also contains information about the type of variable and its categories, when applicable.

Table 2.1: Description of the variables included on data set.

Variable	Description	Type	Categories (when applicable)
ID	Patient identifier - anonymised identification number	Character	
age	Patient age at diagnosis	Quantitative Continuous	
gender	Patient gender	Categorical (Nominal)	Female, Male
district	District of residence	Categorical (Nominal)	Beja, Évora, Faro, Ilha da Madeira, Leiria, Lisboa, Portalegre, Santarém, Setúbal
county	County of residence	Categorical (Nominal)	Abrantes, Alandroal, Albufeira, Alcácer do Sal, Alcanena, Alcobaça, Alcochete, Alenquer, Aljezur, Aljustrel, Almada, Almeirim, Almodôvar, Alpiarça, Alter do Chão, Alvito, Amadora, Arraiolos, Arronches, Arruda dos Vinhos, Avis, Azambuja, Barreiro, Beja, Benavente, Bombarral, Borba, Cadaval, Caldas da Rainha, Calheta, Câmara dos Lobos, Campo Maior, Cartaxo, Cascais, Castro Marim, Castro Verde, Chamusca, Constância, Coruche, Crato, Elvas, Entroncamento, Estremoz, Évora, Faro, Ferreira do Algarve, Ferreira do Zêzere, Fronteira, Funchal, Gavião, Golegã, Grândola, Lagoa, Lagos, Lisboa, Loulé, Loures, Lourinhã, Mação, Machico, Mafra, Marvão, Mértola, Moita, Monchique, Montemor-O-Novo, Montijo, Mora, Moura, Mourão, Nazaré, Nisa, Óbidos, Odemira, Odivelas, Oeiras, Olhão, Ourém, Ourique, Palmela, Peniche, Ponte de Sôr, Portalegre, Portel, Portimão, Redondo, Reguengos de Monsaraz, Ribeira Brava, Rio Maior, Salvaterra de Magos, Santa Cruz, Santana, Santarém, Santiago do Cacém, São Brás de Alportel, Sardoal, Seixal, Serpa, Sesimbra, Setúbal, Silves, Sines, Sintra, Sobral de Monte Agraço, Tavira, Tomar, Torres Novas, Torres Vedras, Vendas Novas, Viana do Algarve, Vidigueira, Vila do Bispo, Vila Franca de Xira, Vila Nova da Barquinha, Vila Real de Santo António, Vila Viçosa
top	Primary tumor site	Categorical (Nominal)	C34.0, C34.1, C34.2, C34.3, C34.4, C34.5, C34.6, C34.7, C34.8, C34.9
mor	Tumor morphology	Categorical (Nominal)	M8012/3, M8046/3, M8070/3, M8071/3, M8072/3, M8073/3, M8140/3, M8230/3, M8250/3, M8251/3, M8252/3, M8254/3, M8255/3, M8260/3, M8480/3, M8560/3, M8550/3
laterality	Laterality of the tumor	Categorical (Nominal)	Left, Right
stage_diag	Stage at diagnosis	Categorical (Ordinal)	I, II, IIIA, IIIB, IV
status	Vital Status	Categorical (Nominal)	Dead, Alive

Some variables had missing values, such as stage at diagnosis - dependent variable, and laterality - independent variable. Additionally, laterality was sometimes registered as "Bilateral", which is considered mal-practice according to registry rules. Thus, patients with laterality "Bilateral" and with missing values were excluded from the dataset.

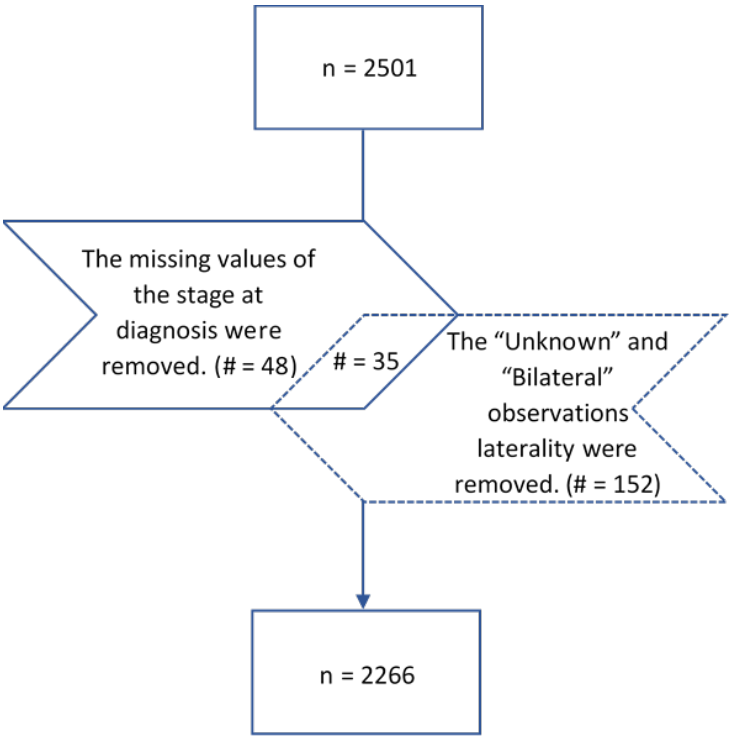

Figure 2.2: Sample size at different phases of data treatment.

2.2 Socioeconomic and access to healthcare data

As seen in the table 2.1, the dataset does not include information about socioeconomic conditions of the patients, nor about access to healthcare. However, this information is crucial to accomplish the main goal of this project - study the association between stage at diagnosis and socioeconomic conditions and access to healthcare. Thus, socioeconomic and access to healthcare indicators were extracted from INE and PORDATA, considering the residence county of each patient.

The socioeconomic conditions variables were selected based on the literature previously described in section 1.3 of the introduction. Hence, the socieconomic factors downloaded were the following:

- monthly average base wage (€) by geographic location (NUTS II), gender and age group;
- declared gross income per inhabitant (€) by county;
- social security pensioners per 1000 inhabitants in active age (‰) by residence county;
- number of welfare recipients per 1000 inhabitants in active age (‰) by residence county

Regarding access to healthcare, the following variables were selected:

- number of nurses per 1000 inhabitants by work county;
- number of medical doctors per 1000 inhabitants by residence county;
- number of official clinics by county and type of service;
- number of attendances at emergency services in hospitals by county

Table 2.2: Description of the socioeconomic indicators.

Variables	Definition	Available data for the years:
dgi	Annual declared gross income per inhabitant (€)	2015-2017
mabs	Monthly average base wage (€)	2007-2017
pens	Social security pensioners per 1000 inhabitants (‰)	2011-2018
benef	Welfare recipients per 1000 inhabitants (‰)	2011-2018
nurs1000	Number of nurses per 1000 inhabitants	2011-2017
doc1000	Number of medical doctors per 1000 inhabitants	2011-2018
numb.oc	Number of official clinics	2011-2012
att	Number of attendances in health centers	2011-2012

As seen from table 2.2, most indicators are available for the period of the study - 2013 and 2014. However, for some indicators, data was not available for the desired period. Thus, the closest year to the diagnosis was considered. Excluding the variable mabs, all indicators can be merged with the main dataset using the residence county. The monthly average base wage is available for different dimensions: NUTS II, gender and age group. Although the main dataset does not have the NUTS II variable, it is possible to identify the respectiveNUTS II region using the residence county, thus adding a new variable in the dataset NUTS. Similarly, a new variable was created - age_cat with the following categories to match with the variable mabs: 18-24; 25-34; 35-44; 45-54; 55-64 and 65+. Thus, using these two new variables and gender, mabs was merged with the main dataset.

The variable pens corresponds to the number of social security pensioners per 1000 inhabitants inactive age by residence county. This indicator was used as a proxy for the social context in which the patient lives. The variable benef corresponds to the number of welfare recipients of the social integration income, of social security by residence county. This integration income is a benefit included in the solidarity subsystem, in order to provide people and their households with support adapted to their personal situation, which contributes to the satisfaction of their essential needs and which favor the progressive insertion of work, social and community. As the previous variable, the welfare recipients help us to define the social context of each patient.

The variable nurs1000 corresponds to the number of available nurses per 1000 inhabitants by work county, and the variable doc1000 corresponds to the number of available doctors per 1000 inhabitants by residence county. The variable numb.oc refers to the number of public health establishments, which aims to promote health, prevent disease and provide care, either intervening in the first line of action of the National Health Service, or ensuring continuity of care, whenever there is a need to use other specialized services and care. It directs its action both to individual and family health and to the health of groups and the community. It may include inpatient service. The variable att refers to the number of medical appointments at health centers by county. The medical appointment is defined by an act of assistance provided by a doctor to an individual consisting of clinical observation, diagnosis, therapeutic prescription, counselling or verification of the evolution of their health status. These four variables serve as a proxy for accessibility to healthcare. However, the variable numb.oc and att are not being weighed by the population of the respective county. So, in order to eliminate the population scale factor, the ratio between each indicator and the resident population by county is considered.

Let's consider the total number of residents in $c-th$ county and $t-th$ year, where $c = 1, ..., 116$ and $t = 1, 2$. Notice that the indexed years correspond to the years of diagnosis - 2013 and 2014.

$$att1000_{ct} = \frac{att_{ct}}{pop_{ct}} \times 1000, \qquad c = 1, 2, 3, ..., 116, \;\; t = 1, 2 \tag{2.1}$$

$$numb.oc1000_{ct} = \frac{numb.oc_{ct}}{pop_{ct}} \times 1000, \qquad c = 1,2,3,...,116, \;\; t = 1,2 \tag{2.2}$$

So, $att1000_{ct}$ and $numb.oc1000_{ct}$ correspond to the number of attendance per 1000 inhabitants and to the number of official clinics per 1000 inhabitants, respectively, for the $c-th$ county and $t-th$ year.

As seen from table 2.2, some indicators do not have available information for the years of diagnosis. And in these cases, as mentioned previously, the closest period containing information was considered. However, there is one indicator - mabs - in which the information is not available for one region during this period, Autonomous Region of Madeira. It has information in 2013 and 2014 for Center, Alentejo, Algarve and Lisbon Metropolitan Area, but for Autonomous Region of Madeira, the last period with available data is 2009. Thus, the monthly average base salary was compared between the Autonomous Region of Madeira and the others for 2009. On average, the monthly wage is 727€ in Alentejo, 747€ in Algarve, 1056€ in Lisbon Metropolitan Area, 745€ in Center and 858€ in Autonomous Region of Madeira. With a variation rate of -12.9%, Algarve is the most similar region to Madeira. So, the assumption that Madeira evolved in the same way as the Algarve is assumed.

Let's consider that Δ_{rtag} corresponds to the relative variation in $r-th$ NUTS II regions between 2009 and the $t-th$ year, considering the $a-th$ age group and $g-th$ gender. Regarding monthly average base wage, if Algarve is the most similar region when compared to Autonomous Region of Madeira in 2009, the variation for this region is given by:

$$\Delta_{rtag} = \frac{mabs_{r2009ag} - mabs_{rtag}}{mabs_{rtag}}, \qquad r = 1,2,...5, \; t = 1,2, \; a = 1,...,5, \; g = 1,2 \tag{2.3}$$

Table 2.3 displays the monthly average base wage of a person who lives in Algarve, according to his age, gender and years of diagnosis. The sixth column is used in order to rebuild the monthly average base wage for Autonomous Region of Madeira.

Table 2.3: Estimated of monthly average base wage for Autonomous Region of Madeira in the years of diagnosis.

year	age_cat	gender	$mabs_{Algarve2009ag}$	$mabs_{Algarvetag}$	$\Delta_{Algarvetag}$	$mabs_{ARM2009ag}$	$mabs_{ARMtag}$
2013	18-24	Male	583.38	633.11	0.085	580.33	629.80
	25-34		749.37	739.08	-0.014	777.39	766.72
	35-44		8868.15	73.033	0.006	954.73	960.10
	45-54		911.72	925.55	0.015	1060.88	1076.97
	55-64		910.3	964.72	0.06	1108.83	1175.12
	65+		863.09	923.28	0.07	1154.08	1234.56
	18-24	Female	575.42	577.73	0.004	564.18	566.44
	25-34		709.71	711.63	0.003	746.8	748.82
	35-44		725.99	767.29	0.057	762.07	805.42
	45-54		703.8	735.56	0.045	768.43	803.11
	55-64		678.59	718.22	0.058	788.66	834.72
	65+		684.42	731.22	0.068	1029.87	1100.29
2014	18-24	Male	583.38	617.41	0.058	580.33	614.18
	25-34		749.37	728.21	-0.028	777.39	755.44
	35-44		868.15	857.92	-0.012	954.73	943.48
	45-54		911.72	911.02	-0.001	1060.88	1060.07
	55-64		910.3	945.44	0.039	1108.83	1151.63
	65+		863.09	887.60	0.028	1154.08	1186.85
	18-24	Female	575.42	584.58	0.016	564.18	573.16
	25-34		709.71	711.98	0.003	746.8	749.19
	35-44		725.99	770.60	0.061	762.07	808.90
	45-54		703.8	741.17	0.053	768.43	809.23
	55-64		678.59	723.45	0.066	788.66	840.80
	65+		684.42	736.93	0.077	1029.87	1108.88

For the Autonomous Region of Madeira, the monthly average base wage is estimated for 2013 and 2014, based on its value in 2009, compounded by the variation rate, $\Delta_{Alrgarvetag}$.

$$mabs_{ARMtag} = mabs_{ARM2009ag} \times (1 + \Delta_{Algarvetag}), \quad t = 1,2, \;\; a = 1,2,...,5, \;\; g = 1,2 \tag{2.4}$$

This variable combines several features (age group, gender and NUTS II), which allows an greater variability and differentiation when compared to an indicator whose aggregation level is only residence region. But the residence region of variable mabs is a macro view (NUTS II). However, there is an additional indicator in the data set, in which the patient is characterized about his economic situation, dgi, where it depends exclusively on the residence county. So, combining these two indicators - mabs and dgi - it is possible to build more accurate and individual information regarding patients' income.

In order to combine these indicators, we started by standardizing the variable mabs. Thus, for a patient diagnosed at $t-th$ year with $g-th$ gender who lives in $r-th$ region, that belongs to the $a-th$ age group is attributed a rank, f_{rtag}, ranging from 0 to 1.

$$f_{rtag} = \frac{mabs_{rtag} - \min(mabs_{rt})}{\max(mabs_{rt}) - \min(mabs_{rt})}, \quad r = 1,2,...,5, \;\; t = 1,2, \;\; a = 1,2,...,5, \;\; g = 1,2 \tag{2.5}$$

Both counties and NUTS II are two indicators which identify the residence location. The difference between these two is that the residence location is more detailed with counties than with NUTS II. In fact, the counties are nesting to the NUTS II. So, the equation 2.5 can be rewritten as follows:

$$f_{ctag} = \frac{mabs_{ctag} - \min(mabs_{ct})}{\max(mabs_{ct}) - \min(mabs_{ct})}, \quad c = 1,2,...,116, \;\; t = 1,2, \;\; a = 1,2,...,5, \;\; g = 1,2 \tag{2.6}$$

Then, the product between f_{ctag} and dgi_c gives rise to a new variable, called *income.index*.

$$income.index_{ctag} = f_{ctag} \times dgi_c, \quad c = 1,2,...,116, \;\; t = 1,2, \;\; a = 1,2,...,5, \;\; g = 1,2 \tag{2.7}$$

So, this new variable is used to evaluate the economic condition of each patient, instead of using dgi and mabs. In order to decrease the order magnitude, a transformation into thousands was applied to the estimated annual income.

Some of the variables considered might be correlated to each other, for instance, the number of attendances (att) might grow with the number of official clinics (numb.oc). Hence, the presence of multicollinearity was assessed. Multicollinearity is a problem in which one independent variable is strongly linearly correlated with one or more independent variables. In this situation, a unique least-squares solution for regression coefficients does not exist and the marginal contribution of that independent variable is influenced by other independent variables. Thus, the linear correlation between the independent variables set should be analyzed.

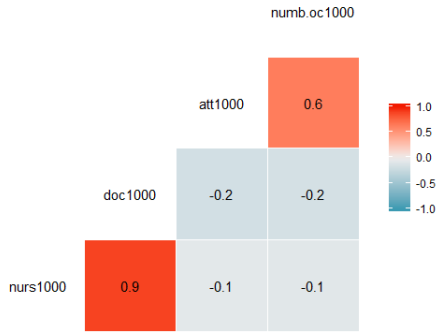

Figure 2.3: Correlation coefficient between access to healthcare indicators.

Figure 2.4: Correlation coefficient between socioeconomic condition indicators.

Figure 2.3, presents the correlations coefficients between access to healthcare indicators. The number of nurses per 1000 inhabitants is highly correlated with the number of doctors per 1000 inhabitant ($r = 0.9$), and there is also a positive correlation between the number of official clinics per 1000 inhabitants and the number of attendances per 1000 inhabitants ($r = 0.6$). Apart from that, there are no other strong linear correlations. Thus, the number of doctors per 1000 inhabitants was the chosen indicator for health professionals since doctors make the final diagnosis. The number of attendances per 1000 inhabitants was chosen over the number of official clinics per 1000 inhabitants since the former can be more informative. A higher number of clinics does not assure that all of them are available. On the other hand, the number of attendances is an indicator that represents the healthcare response.

Figure 2.4 presents the correlation coefficients between socioeconomic conditions indicators. The strongest correlation is between the number of welfare recipients per 1000 inhabitants and the number of pensioners per 1000 inhabitants, $r = 0.4$. Welfare recipients are people in extreme poverty who receive monetary support [1]. Thus, counties where this indicator is higher correspond to the poorest ones, and for that reason that indicator was chosen over the number of pensioners.

Figure 2.5 presents the variables included in the final dataset.

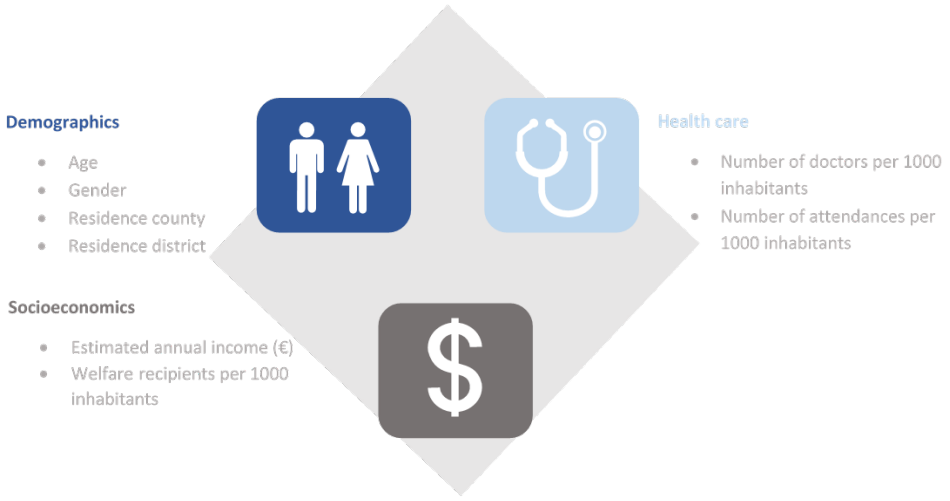

Figure 2.5: Demographic, socioeconomic and access to healthcare variables available.

Chapter 3

Methodology

In the present chapter the applied models to assess the impact of each variable will be presented. The models' choice is justified not only by the typology of the response variable but also by the characteristics of the data collection and by the research question. As explained in chapter 2 the variable of interest is of the ordinal categorical type with five categories.

Since the response variable - stage at diagnosis - is ordinal, the most appropriate model is the ordinal logistic regression, which is an extension of the multinomial logistic regression. There are three variants of the ordinal logistic regression model with respect to the imposition or relaxation of constraints involving the linear systematic component of the model - proportional odds case, partial proportional odds case or unconstrained case. The first one is the most restrictive but also the most parsimonious case, where the parallel lines assumption holds for all outcome categories, that is, the same coefficient vector for all outcome categories is assumed.

The partial proportional odds case relaxes the parallel lines assumption for a subset of the regression coefficients across the outcome categories. The unconstrained case is the most flexible and least parsimonious case. It relaxes the parallel lines assumption for all outcome categories, where the difference of the regression coefficients for each outcome category is allowed. These assumptions will be assessed in order to choose the most adequate approach.

The assumption of independence of observations applies to all regression models, but there is a particularity in the data of the current project. All the variables regarding socioeconomic conditions are collected at the residence county, which means that patients living at the same county have the same information. One of the approaches used to handle clustered or grouped data is to consider a mixed model, where random and fixed are included.

Independence test

The chi-square (χ^2) independence test is used to analyze the joint behaviour or two categorical variables. It evaluates whether there is a significant association between the categories of the two variables. The building of χ^2 test is based on following hypotheses:

H_0: There is no association between categories of the two variables

H_1: The categories of the two variables are associated

Test statistic:

$$X^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-e_{ij})^2}{e_{ij}} \underset{H_0}{\sim} \chi^2_{((r-1)(c-1))} \tag{3.1}$$

where r and c correspond to the number of rows and columns in the contingency table, respectively. In equation (3.1), O_{ij} represents the observed frequency of category i of the first variable, and category j of the second variable, simultaneously, and e_{ij} is the expected frequency, where $O_{i\cdot}$ are the marginal observations of category i of the first variable and $O_{\cdot j}$ are the marginal observations of category j of the second variable.

$$e_{ij} = \frac{\sum_{i=1}^{r} O_{i\cdot} \times \sum_{j=1}^{c} O_{\cdot j}}{N}, \tag{3.2}$$

where N is the total number of observations.

The hypothesis H_0 is rejected if $X^2_{obs} > \chi^2_{(1-\alpha)}((r-1)(c-1))$ or, the *p-value* associated to the test statistic is smaller than α, the significance level. *p-value* is the probability of observing a sample statistic as extreme or more extreme than the statistic test: $P(\chi^2_{(r-1)(c-1)} \geq X^2_{obs})$.

3.1 Multinomial Logistic Regression

Multinomial logistic regression is a widely used statistical modeling technique. This procedure is an extension of the binary logistic regression.

In multinomial logistic regression, the response variable is a qualitative nominal variable with more than two categories. The independent variables can be either qualitative (factor) or quantitative (numeric).

Consider the response variable $Y_i, \forall i = 1, ..., n$, which can be equal to one of the several categories labelled j, with $j = 1, ..., J$, and let $\pi_{ij} = P(Y_i = j|\mathbf{x}_i)$ correspond to the probability that the $i-th$ response falls in the $j-th$ category. Since the J categories cover all possibilities and are mutually exclusive, $\sum_{j=1}^{J} \pi_{ij} = 1, \forall i$. The number of groups corresponds to the number of categories of the response variable (J groups). There are J different groups of different sizes. Let n_j denote the number of cases in the $j-th$ group. Then, $\sum_{j=1}^{J} n_j = n$, that is, the sum of sizes of each group is the sample size.

Let Y_{ij} be the number of responses from the $i-th$ group that fall in the $j-th$ category. With observed value y_{ij}, the probability distribution of $Y_{ij}, \forall j = 1, ..., J$ is given by the multinomial distribution:

$$P(Y_{i1} = y_{i1}, ..., Y_{iJ} = y_{iJ}) = \binom{n_i}{y_{i1}...y_{iJ}} \pi_{i1}^{y_{i1}} ... \pi_{iJ}^{y_{iJ}}, \qquad i = 1, ..., n \tag{3.3}$$

Consider n independent observations, p explanatory variables and a response variable has J categories. In multinomial data the most appropriate approach is to consider one of the response categories as a baseline or reference value.

The *logit* transformation of π_{ij} is applied [13]. This transformation corresponds to consider the logarithm of the odds of category j, taking category J as the reference. The importance of this transformation is that the function has the desirable proprieties of a linear regression model. It is linear in its variables which may be continuous and may range from $-\infty$ to $+\infty$, depending on the range of the independent variables. So, in terms of π_{ij}, the $logit$ transformation is defined as:

$$logit(\pi_{ij}) = log\left(\frac{\pi_{ij}}{\pi_{iJ}}\right) = log\left(\frac{\pi_{ij}}{1 - \sum_{i=1}^{J-1} \pi_{ij}}\right) = \beta_{0j} + \beta_{1j}x_{i1} + \beta_{2j}x_{i2} + ... + \beta_{pj}x_{ip}, \tag{3.4}$$

where $(\beta_{0j}, \beta_{1j}, ..., \beta_{pj})$ are the regression coefficients, with $j = 1, ..., J$ and p corresponds to the number of independent variables. Each regression coefficient refers to the effect of the associated variable on the $log(odds)$. The unknown parameters $(\beta_{0j}, \beta_{1j}, ..., \beta_{pj})$ are estimated through maximum likelihood.

Summing up, the logistic regression analyses for categorical outcomes attempts to model the odds of

event's occurrence and to estimate the effects of independent variables on these odds. These measures are a ratio comparing the probability that an event occurs (referred as "success") with the probability that it does not occur (referred as "failure"). When the probability of success is greater than the probability of failure, the odds is greater than 1. However, if the two outcomes are equal, the odds are 1. When the probability of success is lower than the probability of failure, the odds is less than 1.

To evaluate the impact on the odds of an independent variable, OR are calculated, which compares the odds for different values of the explanatory variable. OR are bounded below by 0 but have no upper bound. So, OR is a measure of association between the binary outcome and an independent variable that provides "a clear indication of how the risk of the outcome being present changes with the variable in question" [13].

3.2 Ordinal Logistic Regression

When the possible responses for an outcome variable consist in more than two ordered categories, the most suitable model is the ordinal logistic regression model [4].

The complexity in fitting the ordinal regression models arises in part because there are many different possibilities for how "success" might be modeled. Generally, success corresponds to an event of interest, but as there are more than two possibilities success becomes a relative term. This analysis is referred to as cumulative odds. It is one way to conceptualize how the data might be sequentially partitioned into dichotomous groups, while still taking advantage of the order of the response categories. The cumulative *logit* model was proposed by Walker and Duncan [38] and later called the proportional odds model by McCullagh [19].

3.2.1 The model

Consider a multinomial response variable Y with categorical outcomes, denoted by $1, 2, \ldots, J$ and $\mathbf{x}_i$, $i = 1, ..., n$ denote a $p - dimensional$ vector of explanatory variables. The dependence of Y on X for the proportional odds model is the following:

$$\begin{aligned} logit(\pi_{ij}) = ln\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) &= ln\left(\frac{P(Y_i \leq j|\mathbf{x}_i)}{1-P(Y_i \leq j|\mathbf{x}_i)}\right) \\ &= ln\left(\frac{P(Y_i \leq j|\mathbf{x}_i)}{P(Y_i > j|\mathbf{x}_i)}\right) \\ &= \alpha_j + \beta_1 x_1 + ... + \beta_p x_p, \end{aligned} \tag{3.5}$$

where $\pi_{ij} = P(Y_i \leq j|\mathbf{x}_i)$, where $j = 1, ..., J-1$ represents the probability that a response falls in a category less or equal to the $j-th$ category for the since $i-th$ patient, that is, $Y_i \leq j$. J is the total number of categories, the cumulative probability of J categories is 1, $P(Y_i \leq J) = 1$. So, the information about this category is redundant. The parameters α_j, with $j = 1, ..., J-1$, are the unknown intercept parameters, satisfying the condition $\alpha_1 \leq \alpha_2 \leq ... \leq \alpha_{J-1}$. The α_j terms, called threshold values, correspond to the intercept in a linear regression, but the difference is that, in the ordinal logistic regression model, the number of thresholds correspond to the number of categories of the response variable. $\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_p)'$ is the vector of unknown regression coefficients.

The ordinal logistic regression model is parameterized as

$$logit(\pi_{ij}) = \alpha_j - (\beta_1 x_{i1} + ... + \beta_p x_{ip}), \qquad i = 1, ..., n \;\; j = 1, ..., J-1 \tag{3.6}$$

The cumulative *logits* associated with being at or below a particular category j can be exponentiated to obtain estimated cumulative odds and then used to find the estimated cumulative probabilities associated with being at or below category j.

In the model expression (3.7) there is a minus sign before the coefficients for the predictor variables, instead of the customary plus sign. That is done so that larger coefficients reveal an association with larger scores. Usually, a positive coefficient for a dichotomous factor is associated to higher scores and these ones are more likely for the first category. A negative coefficient means that lower scores are more likely. For a continuous variable, a positive coefficient means that with the increase of the variable, the likelihood of larger scores increases. An association with higher scores means smaller cumulative probabilities for lower scores, since they are less likely to occur.

$$\ln\left(\frac{P(Y_i \leq j|\mathbf{x}_i)}{P(Y_i > j|\mathbf{x}_i)}\right) = \alpha_j - (\beta_1 x_{i1} + ... + \beta_p x_{ip}) \Leftrightarrow \frac{P(Y_i \leq j|\mathbf{x}_i)}{P(Y_i > j|\mathbf{x}_i)} = \exp(\alpha_j - (\beta_1 x_{i1} + ... + \beta_p x_{ip})) \tag{3.7}$$

$$i = 1, ..., n, \;\; j = 1, ..., J-1$$

The regression coefficients, $\boldsymbol{\beta}$, do not depend on j, implying that the model assumes that the relationship between x_l, $l = 1, ..., p$, and Y is independent of j. That is, it implies that the explanatory variables have the same effect on the odds, regardless the different consecutive 'splits' to the data for each category of the model. So, each *logit* has its own α term but the same coefficients $\boldsymbol{\beta}$. McCullagh [19] calls this assumption of identical *log-odds ratios* across the $(J-1)$ cut points the proportional odds assumption, and hence the name proportional odds model. This assumption can be checked based on a chi-square score test [27].

A model that relaxes the proportional odds assumption can be represented as $logit(\pi_{ij})$ where the regression parameter vector is allowed to vary with j. Likelihood ratio test or Wald Chi-Square test are used to test this hypothesis (Long,1997; Agresti, 2002). In ordinal *logit* regression, this test examines the equality of the different categories, $H_0 : \beta_{1j} = \beta_{2j} = \ldots = \beta_{pj}$, $j = 1, ..., J$. The rejection of the assumption of parallelism for the particular ordinal model being investigated implies that, at least, one of the explanatory variables may have a different effect across the outcome levels, that is, there is an interaction between one or more independent variables and the splits.

When the assumption is not verified, the proportional odds model is not valid. Alternatively, there is a flexible model that accommodates variables whose effect is different depending on the category of the response variable. Suggested by Peterson and Harrell [27], partial proportional odds model can be used when parallel lines assumption is not valid. This model is known as partial proportional odds model due to the fact that only a few variables respect the proportional odds assumption. Hence, this model contains variables whose effect varies with the category of the response variable - violate the assumption - and variables whose effect remain despite the category of the response variable.

The expression is the following:

$$\ln\left(\frac{P(Y_i \leq j|\mathbf{x}_i)}{P(Y_i > j|\mathbf{x}_i)}\right) = \alpha_j + \beta_{1j}x_{i1} + ... + \beta_{mj}x_{im} - \beta_{(m+1)}x_{i(m+1)} - ... - \beta_p x_{ip}$$

$$\Leftrightarrow \frac{P(Y_i \leq j|\mathbf{x}_i)}{P(Y_i > j|\mathbf{x}_i)} = \exp(\alpha_j + \beta_{1j}x_{i1} + ... + \beta_{mj}x_{im} - \beta_{m+1}x_{i(m+1)} - ... - \beta_p x_{ip}) \tag{3.8}$$

where $i = 1, ..., n, \;\; j = 1, ..., J-1$. There are m variables for which the proportional odds assumption is not verified and there are $p-m$ variables for which this assumption is verified. Thus, $m \times (j-1)$ coefficients are estimated for vector $(x_1, ..., x_m)$ and $p-(m+1)$ are estimated for vector $(x_{m+1}, ..., x_p)$, meaning that the model are includes $m \times (j-1) + p - (m+1)$ coefficients.

The estimates of the coefficients in logistic regression are found via maximum-likelihood estimation. To test whether a regression coefficient is significant, the hypothesis that this parameter differs from zero is evaluated. From the asymptotic theory of maximum-likelihood estimation, the parameters' estimators will be approximately normally distributed with null mean, assuming that the hypothesis null is true. The standard error of a statistic is the approximate standard deviation of a statistical sample population. It provides the absolute measure of the distance that the data points (sample) fall from the regression line (predicted values). So, the hypotheses to be tested are

$$H_0 : \beta_k = 0 \qquad vs. \qquad H_1 : \beta_k \neq 0, \qquad k = 1, ..., p$$

and the Wald statistic is defined as

$$W = \frac{(\hat{\beta_k} - \beta_{k|H_0})}{\hat{se}(\hat{\beta_k})} \sim N(0,1) \tag{3.9}$$

Under the null hypothesis, that is, if $\beta_{k|H_0} = 0$, the Wald statistics simplifies to:

$$W = \frac{\hat{\beta_k}}{\hat{se}(\hat{\beta_k})} \sim N(0,1) \tag{3.10}$$

3.2.2 Odds Ratio

The OR is a measure of association between an exposure variable and the outcome. It represents the odds that an outcome will occur given a particular exposure, compared to the odds of the outcome occurring in the absence of that exposure.

In logistic regression, the regression coefficients correspond to the estimated increase in the *log odds* of the outcome (dependent variable) per unit increase in the value of the exposure (independent variable), if the independent variable is quantitative. That is, the exponential function of the regression coefficient is the OR associated increasing one unit in the exposure. When the independent variable is binary, the exponential function of its regression coefficient is the OR associated to the presence of the variable in the outcome against to its absence. In the proportional odds model, the OR are calculated in the same way, but only if the event of interest is $Y \leq j$. If $Y > j$ is the intended, some transformations are required. A commonly used link function is the *logit*, which leads to

$$logit(P(Y_i \leq j)) = \ln\left(\frac{P(Y_i \leq j)}{1 - P(Y_i \leq j)}\right) \qquad i = 1, ..., n, \;\; j = 1, ..., J-1 \tag{3.11}$$

Consider the binary variable x where $x = 1$ represents exposure and $x = 0$ represents the absence of exposure. In this scenario, zero corresponds to the baseline or reference class. Thus, the OR of the event $Y \leq j$ at $x = 0$ relative to the same event at $x = 1$ is given by

$$OR = \left(\frac{odds_{x=1}}{odds_{x=0}}\right) = \frac{\left(\frac{P(Y \leq j|x=1)}{1-P(Y \leq j|x=1)}\right)}{\left(\frac{P(Y \leq j|x=0)}{1-P(Y \leq j|x=0)}\right)} = \frac{\exp(\beta_0 - \beta_1)}{\exp(\beta_0)} = \exp(-\beta_1) \tag{3.12}$$

However, sometimes in this type of model it could be interesting to assess this value for the opposite event, $Y > j$. In this case, the solution is to reverse expression (3.12), that is,

$$\begin{aligned} OR = \left(\frac{odds_{x=1}}{odds_{x=0}}\right) &= \frac{\left(\frac{1-P(Y \leq j|x=1)}{P(Y \leq j|x=1)}\right)}{\left(\frac{1-P(Y \leq j|x=0)}{P(Y \leq j|x=0)}\right)} \\ &= \frac{\left(\frac{1}{\exp(\beta_0 - \beta_1)}\right)}{\left(\frac{1}{\exp(\beta_0)}\right)} = \exp(\beta_1) \end{aligned} \tag{3.13}$$

As seen in equations (3.12) and (3.13), OR does not depends on j. Thus, cumulative OR is proportional to the distance between the x values. In equation (3.8), the value of OR changes according to each category j for the variables whose assumption of parallel lines is not verified.

Table 3.1: Impact of independent variables through the OR value

OR = 1	Exposure does not affect odds of outcome
OR >1	Exposure associated with higher odds of outcome
OR <1	Exposure associated with lower odds of outcome

In the appendix section, a practical example of the proportional odds model and a practical example of the partial proportional odds model are presented. From these, the interpretation of regression parameters is made as well as the OR.

3.3 Ordinal Generalized Linear Mixed Models

The generalized linear mixed model (GLMM) is an extension of the generalized linear model that includes random effects as well as fixed effects in the linear predictor. It is important to note that GLMMs for ordinal responses assume multinomial distribution at each particular value of the fixed and random effects.

The random effect provides the correlation expected between observations in the same cluster and allows inference to be made to the population from which the groups were sampled.

Let Y_i, $i = 1, ..., n$, denote $i-th$ observation, consider $j = 1, 2, ..., J$ are the outcome categories and there are p independent variables. The cumulative *logit* model of proportional odds including a random intercept is the following:

$$logit(P(Y_i \leq j)) = \alpha_j + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} + u_i, \tag{3.14}$$

$$i = 1, ..., n, \quad j = 1, ..., J-1$$

The difference between this expression and (3.7) is the addition of the random effect, u_i, to the intercept term α_j. The same random effect for each cumulative probability is assumed, in which the original thresholds $\alpha_1, ..., \alpha_{J-1}$ are simultaneously shifted yielding the thresholds $\alpha_1 + u_i, ..., \alpha_{J-1} + u_i$. Thus, the effective thresholds vary across clusters.

A subject with a relatively large positive/negative u_i has relatively large/small cumulative probabilities, hence a relatively high/low chance of occurring at the low end of the ordinal scale. Thus, the expression of the model can be written alternatively as:

$$logit(P(Y_i \leq j)) = \alpha_j - (u_i + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip}) \tag{3.15}$$

$$i = 1, ..., n, \quad j = 1, ..., J-1$$

Considering this parameterization, high values of random and fixed effects means higher probability of $i-th$ observation coincides with the highest level of the ordinal scale. As in the previous model (3.8), the intercept parameters satisfy $\alpha_1 < \alpha_2 < ... < \alpha_{J-1}$, to reflect the ordering of cumulative probabilities and unlike the explanatory variables parameters, the values of u_i are unknown.

This last model (3.15) displays a cumulative probability of the $i-th$ falling in the $j-th$ category or below where i represents the observations and $j = 1, ..., J$ corresponds to the categories of the response variable. Similarly to the previous models, α_j corresponds to the threshold parameters or cut-points. u_i is a random variable corresponding to the random effects. These random effects are characterized as independent and identically distributed (IID) normal: $u_i \sim N(0, \sigma_u^2)$. The estimate of the variance, $(\hat{\sigma}_u^2)$, describes the variability among clusters, indicating the heterogeneity caused by not including certain explanatory variables that are associated with the response variable.

The practical application of this type of model is displayed in appendix section in order to better understand its interpretation.

Chapter 4

Application

This chapter contains the exploratory analysis, including a characterisation of each variable analysed. Disease mapping of the socioeconomic conditions and the healthcare indicators using the counties of Portugal are presented. The main purpose of this exploratory analysis is to compare the patterns displayed in the maps mentioned above with maps displaying the proportion of individuals in each diagnosis stage. Finally, the cumulative link model and cumulative link mixed model were applied, and the OR was analysed to assess how each variable is related to the severity of non-small lung cancer.

4.1 Exploratory Analysis

The absolute and relative frequencies by cancer stage and overall are used to characterize the qualitative variables. Minimum, median, maximum, mean and standard deviation are the measures used to characterize the quantitative variables by stage and overall.

The association between the qualitative variables and the stage at diagnosis is analyzed through independence tests. Usually, the χ^2 independence test is used to determine if there is a significant association between two categorical variables. According to the literature, the χ^2 test is appropriate when 20% of the expected values is less than 5 at most.

Table 4.1: Summary of the qualitative variables by stage and overall.

Variables	Stage I	Stage II	Stage IIIA	Stage IIIB	Stage IV	Total	p-value
Gender							<0.001 [a]
Female	100 (34.7%)	22 (15.1%)	54 (19.9%)	42 (19.2%)	366 (27.3%)	584 (25.8%)	
Male	188 (65.3%)	124 (84.9%)	217 (80.1%)	177 (80.8%)	976 (72.7%)	1,682 (74.2%)	
District							<0.001[a]
Beja	3 (1.0%)	1 (0.7%)	9 (3.3%)	18 (8.2%)	42 (3.1%)	73 (3.2%)	
Évora	6 (2.1%)	3 (2.1%)	8 (3.0%)	5 (2.3%)	34 (2.5%)	56 (2.5%)	
Faro	24 (8.3%)	12 (8.2%)	23 (8.5%)	31 (14.2%)	167 (12.4%)	257 (11.3%)	
Ilha da Madeira	5 (1.7%)	10 (6.8%)	10 (3.7%)	2 (0.9%)	53 (3.9%)	80 (3.5%)	
Leiria	7 (2.4%)	3 (2.1%)	5 (1.8%)	3 (1.4%)	31 (2.3%)	49 (2.2%)	
Lisboa	188 (65.3%)	77 (52.7%)	144 (53.1%)	104 (47.5%)	658 (49.0%)	1,171 (51.7%)	
Portalegre	4 (1.4%)	6 (4.1%)	6 (2.2%)	9 (4.1%)	17 (1.3%)	42 (1.9%)	
Santarém	14 (4.9%)	11 (7.5%)	19 (7.0%)	18 (8.2%)	117 (8.7%)	179 (7.9%)	
Setúbal	37 (12.8%)	23 (15.8%)	47 (17.3%)	29 (13.2%)	223 (16.6%)	359 (15.8%)	
Laterality							0.821[a]
Left	127 (44.1%)	64 (43.8%)	119 (43.9%)	102 (46.6%)	569 (42.4%)	981 (43.3%)	
Right	161 (55.9%)	82 (56.2%)	152 (56.1%)	117 (53.4%)	773 (57.6%)	1,285 (56.7%)	
Status							<0.001[a]
Dead	87 (30.2%)	87 (59.6%)	209 (77.1%)	196 (89.5%)	1,282 (95.5%)	1,861 (82.1%)	
Alive	201 (69.8%)	59 (40.4%)	62 (22.9%)	23 (10.5%)	60 (4.5%)	405 (17.9%)	

[a] Chi-Square independence test

Table 4.1 provides the descriptive statistics for the qualitative variables by stage at diagnosis and overall. Of the 2266 individuals diagnosed with non-small lung cancer, 1682 (74.2%) are male. More than half of the patients (59.2%) were diagnosed at stage IV, and most women were diagnosed at the most severe stage (63%). More than half of the patients diagnosed with lung cancer are from Lisbon (51.7%), which is expected since Lisbon is the largest district represented in ROR-Sul. Portalegre is the district with a lower number of diagnosed individuals. No significant differences were found regarding the laterality of the tumour. During the follow-up, 82.1% of the patients died. However, the cause of death was absent, hence no inference can be made regarding deaths caused by lung cancer. Within stage I, the proportion of deaths is lower than the proportion of survivors. The situation reverses as the stage at diagnosis increases, with 96% of patients dying in stage IV. More than half of the patients died during the first year after diagnosis (53%).

Table 4.1 displays the associated *p-values* to χ^2 tests and from these it is possible to conclude that the null hypothesis is rejected for gender, district and status indicating that there is statistical evidence that the categories of gender ($p-value < 0.001$), district ($p-value < 0.001$) and status ($p-value < 0.001$) are associated with the stage at diagnosis. The result for variable laterality indicates that there is no association between the stage at diagnosis and the location of the main tumor.

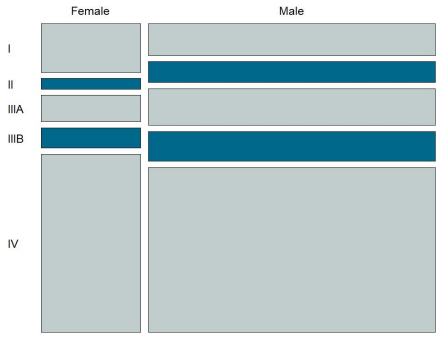

Figure 4.1: Proportion of each stage at diagnosis by gender.

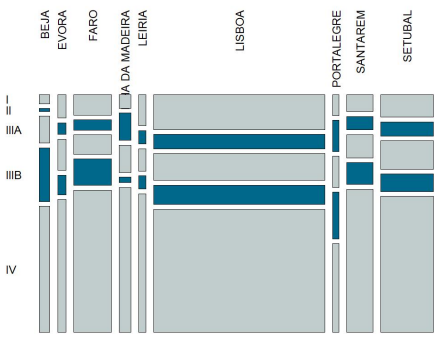

Figure 4.2: Proportion of each stage at diagnosis by district.

Figure 4.3: Proportion of each stage at diagnosis by laterality of the tumour.

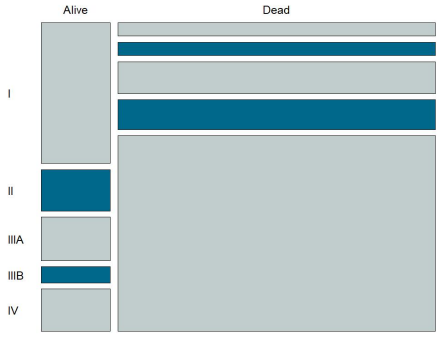

Figure 4.4: Proportion of each stage at diagnosis by status.

The number of male patients is three times higher than the number of female patients (Figure 4.1).

The width of the horizontal bars represents the proportion of female:male indicating that the ratio is about 1:3. The heights of the boxes correspond to the proportion of each stage for females and males. The proportion of female patients in stage I is almost two times higher when compared with the proportion of male patients in stage I. The proportions of females is smaller, compared to males, in stage II, IIIA and IIIB, but higher in stage IV.

Most cases of NSCLC were diagnosed in Lisbon. In Leiria and Portalegre, the proportion of patients diagnosed with NSCLC was lower than the remaining districts. Madeira has a higher proportion of patients diagnosed in stage IV, followed by the districts of Faro and Santarém. The proportion of stage II cancers in Portalegre was the highest and, on the other hand, this proportion was the lowest in Beja. However, when the proportion of stage IIIB cancers was assessed, Beja was the district with the highest value and Madeira presents the lowest proportion of patients diagnosed at stage IIIB cancers. The width of Figure 4.3 indicates that the proportion of patients diagnosed with NSCLC whose primary tumour was located on the right side was slightly higher than cases with a primary tumour on the left side. The height was very similar regardless of the stage at diagnosis.

Figure 4.4 shows that the proportion of patients diagnosed with NSCLC who died during the follow-up period was about four times higher than those who survived. The proportion of patients diagnosed in stage I who survived is much higher when compared with the proportion of patients diagnosed in stage I who died. On the other hand, the proportion of patients diagnosed at stage IV who survived is much lower than those diagnosed at stage IV who died.

The socieconomic as well as the access to healthcare indicators, selected based on the literature review and by the availability of information in PORDATA and INE, were also analyzed. The variable age appeared together with these indicators because all of these were quantitative variables.

As seen previously, education, income, and occupation have been regarded as major potential influencers of health conditions. Hence, as a proxy for healthcare access, the chosen variables were the number of doctors per 1000 inhabitants and the number of attendances per 1000 inhabitants. As a proxy for the socioeconomic conditions of each patient, the chosen variables were the estimated income and the number of social security pensioners per 1000 inhabitants.

Table 4.2: Summary of the quantitative variables by stage at diagnosis and overall.

Variables	Stage I	Stage II	Stage IIIA	Stage IIIB	Stage IV	Total
Age (age)						
min	35	33	31	38	28	28
median	66	67	66	66	65	66
max	88	87	89	88	92	92
mean (sd)	66.27 (9.82)	66.19 (10.74)	65.29 (10.68)	65.37 (10.77)	65.33 (11.37)	65.50 (11.00)
Income (income.index)						
min	1.628853	1.882413	1.492712	1.288244	1.407299	1.288244
median	7.068000	7.424500	7.092353	7.092353	6.710027	6.852066
max	16.655	16.655	16.655	16.655	16.655	16.655
mean (sd)	7.54 (4.02)	8.24 (3.96)	7.81 (3.97)	7.81 (3.78)	7.19 (3.70)	7.44 (3.81)
Welfare recipients (benef)						
min	9.51	9.51	9.71	9.75	8.95	8.95
median	34.86	34.57	31.79	32.92	31.79	32.18
max	120.86	120.86	111.51	122.40	202.89	202.89
mean (sd)	37.56 (15.77)	37.45 (18.60)	36.43 (16.65)	36.83 (18.07)	35.62 (16.08)	36.20 (16.49)
Number of medical doctors (doc1000)						
min	0.4	0.5	0.4	0.4	0.2	0.2
median	3.1	2.9	2.8	2.9	2.8	2.9
max	17.3	17.3	17.3	17.3	17.3	17.3
mean (sd)	6.68 (6.11)	5.88 (5.74)	5.53 (5.54)	5.94 (5.84)	5.21 (5.29)	5.55 (5.53)
Number of attendances (att1000)						
min	1299.9	475.5	1299.9	1299.9	0.0	0
median	2266.92	2266.92	2271.61	2300.00	2300.00	2300
max	6433.77	6433.77	8246.73	6433.77	8307.46	8307.46
mean (sd)	2403.67 (584.73)	2502.30 (962.35)	2495.25 (833.33)	2612.27 (818.85)	2523.94 (879.71)	2512.37 (843.03)

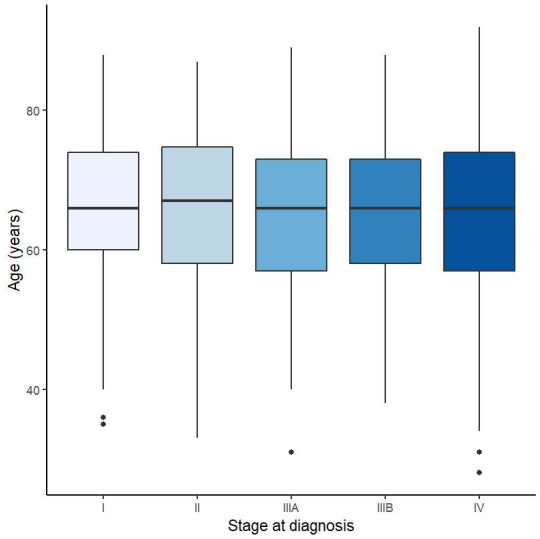

Figure 4.5: Distribution of age by stage at diagnosis.

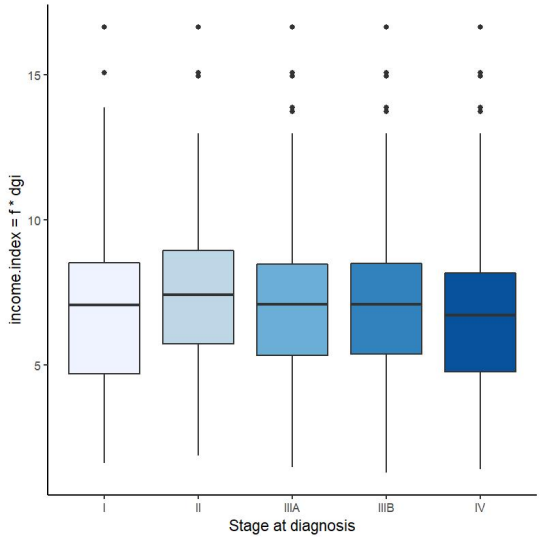

Figure 4.6: Distribution of income by stage at diagnosis.

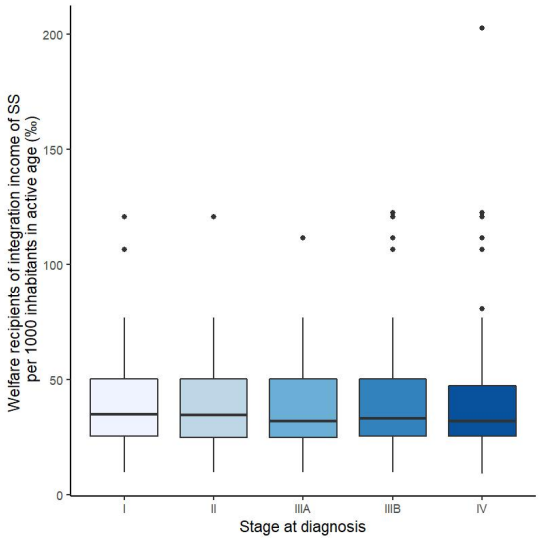

Figure 4.7: Distribution of welfare recipients of integration income of social security per 1000 inhabitants in active age by stage at diagnosis.

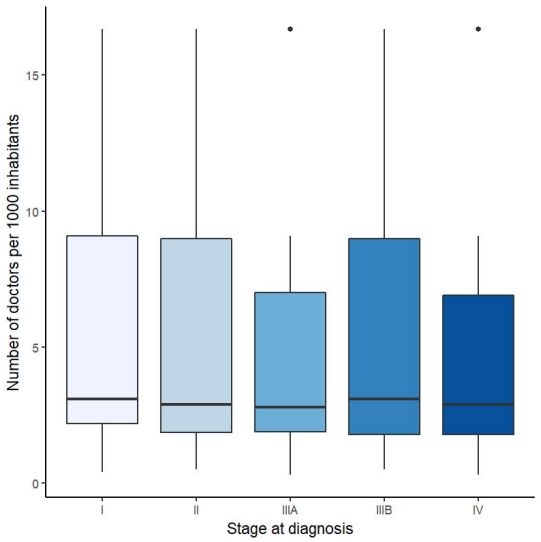

Figure 4.8: Distribution of numbers of doctors per 1000 inhabitants by stage at diagnosis.

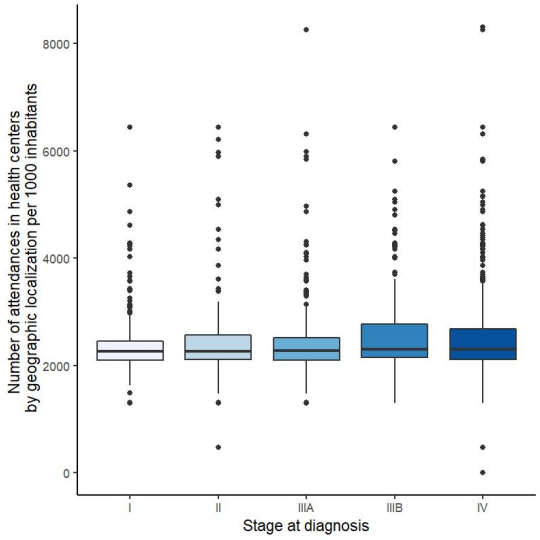

Figure 4.9: Distribution of number of attendances in health centers per 1000 inhabitants by stage at diagnosis.

In Table 4.2, the quantitative variables by stage at diagnosis and overall were summarized with the mean, median, standard deviation, minimum and maximum. The youngest patient diagnosed with NSCLC had 28 years old, and was diagnosed at stage IV. The oldest patient had 92 years old and was also diagnosed at stage IV. The median and the mean age are practically the same between the stages. Figure 4.5 displays the distribution of age between stages. The presence of *outliers* was the main difference between stages I, IIIA and IV and stages II and IIIB. These observations correspond to the patients under 40 years.

As seen previously, the income indicator is an approximation to real annual income. However, it still provides information about the socioeconomic context of each patient. The maximum income was the same for each stage, because there were people with the same characteristics - age, gender and residence location - diagnosed in each stage. But, on average, the estimated annual income was lower for patients

diagnosed at stage IV. The minimum estimated income was observed in people diagnosed at stage IIIB. The highest values corresponded to the *outliers* as shown in Figure 4.6, where the estimated annual income is displayed. Figure 4.7 presents the number of welfare recipients of integration income of social security per 1000 inhabitants. On average, the lowest value was observed in patients diagnosed at stage IV, and the highest value in patients diagnosed at stage I. However, one patient whose residence county has the maximum value of this indicator was diagnosed at the most severe stage, at least.

The maximum number of doctors per 1000 inhabitants was similar across each stage. On average, patients diagnosed in stage I live in counties with a higher number of doctors, 6.68 doctors per 1000 inhabitants. Overall, each county has on average 5.55 doctors per 1000 inhabitants. The distribution of the number of doctors per 1000 inhabitants by stage at diagnosis is displayed in Figure 4.8. The stages I, II and IIIB were equally distributed and the stages IIIA and IV are equally distributed too. Ir is enough that people diagnosed in one of these two groups live in the same county, sharing the value of this indicator. The most important feature for the number of attendances is that the minimum was zero for diagnosed patients at stage IV. However, the maximum value is registered, also, for stage IV. This variable is characterized by having too many *outliers* as seen from the Figure 4.9. Among the patients diagnosed in stage IV, there is one who lives in a context without health care. The lowest number of attendances was registered in a patient who was diagnosed in stage IV and lives in Mourão.

The main purpose of exploratory analysis was to know the data behaviour before making any assumptions. It can help to identify obvious errors, as well as better understand patterns within the data, to detect *outliers* or anomalous events, and to find interesting relations between the variables.

4.2 Disease Mapping

All selected indicators in the section 2 were mapped as well as the proportion of people diagnosed in each stage by residence county. The main purpose of this approach was to visually assess the possibility of associations between these indicators and the severity of lung cancer.

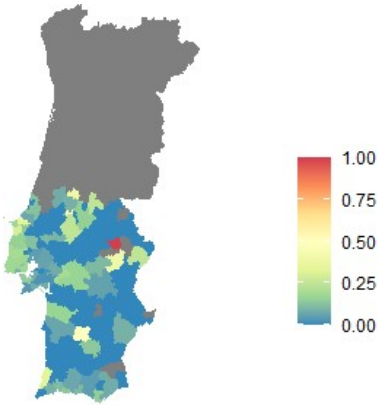

Figure 4.10: Proportion of diagnosed people a stage I for each county of the ROR-Sul area.

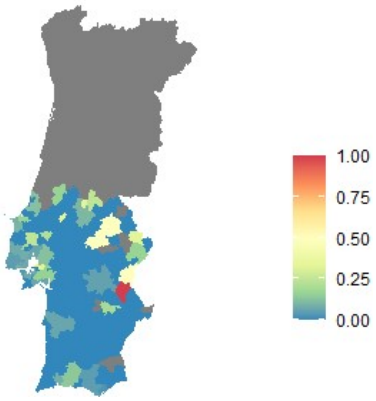

Figure 4.11: Proportion of diagnosed people a stage II for each county of the ROR-Sul area.

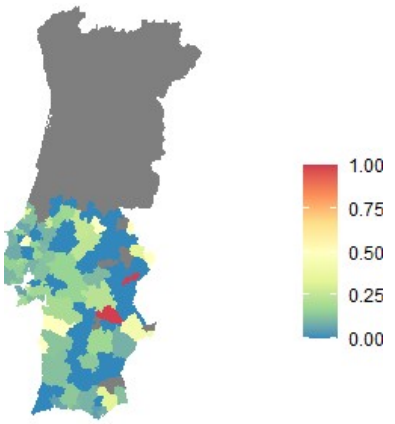

Figure 4.12: Proportion of diagnosed people a stage IIIA for each county of the ROR-Sul area.

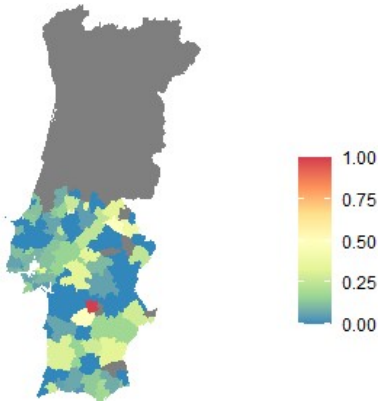

Figure 4.13: Proportion of diagnosed people a stage IIIB for each county of the ROR-Sul area.

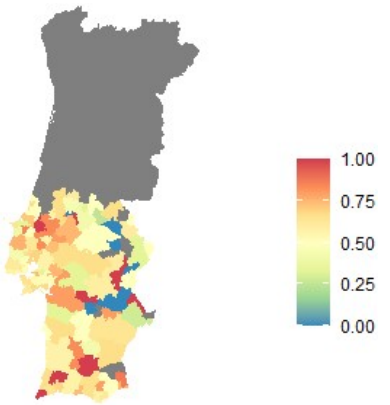

Figure 4.14: Proportion of diagnosed people a stage IV for each county of the ROR-Sul area.

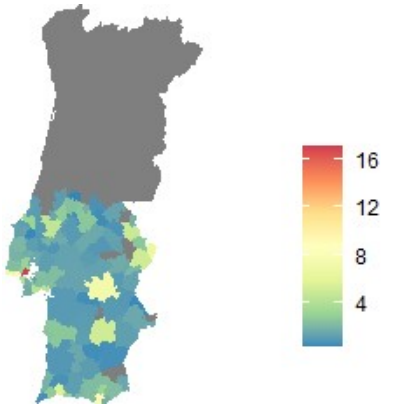

Figure 4.15: Number of doctors per 1000 inhabitants for each county of the ROR-Sul area.

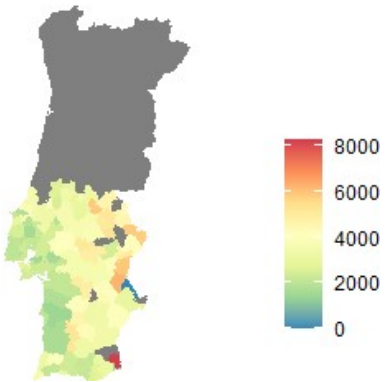

Figure 4.16: Number of attendances per 1000 inhabitants for each county of the ROR-Sul area.

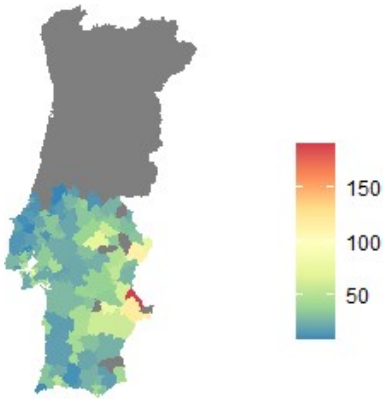

Figure 4.17: Number of welfare recipients per 1000 inhabitants for each county of the ROR-Sul area.

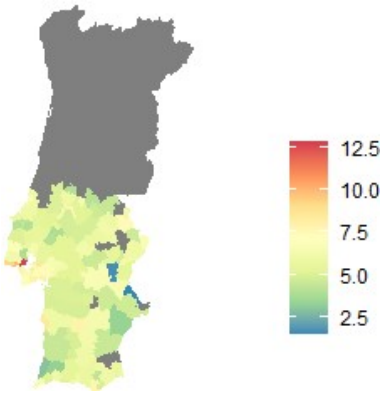

Figure 4.18: Estimated annual income for each county of the ROR-Sul area.

Several counties (Castelo de Vide, Monforte, Sousel, Cuba, Barrancos and Alcoutim) are displayed in grey, corresponding to counties without data. No patients were diagnosed with NSCLC living in those regions. Red counties correspond to higher indicator values, and blue counties correspond to counties with lower indicator values.

According to Figure 4.14, Rio Maior, Constância, Redondo, Borba, Mourão, Viana do Alentejo, Almodôvar, Vila do Bispo and Monchique correspond to the regions with higher proportion of people diagnosed at stage IV. Of this set of counties, Vila do Bispo, Mourão and Viana do Alentejo were included in those with the lowest number of doctors available per 1000 inhabitants. Also, Mourão and Redondo were the counties with the lowest of estimated annual income. Additionally, Mourão was the county with the highest number of welfare recipients per 1000 inhabitants.

Nevertheless, there was no clear pattern between the stage at diagnosis and the socioeconomic and healthcare conditions. The mortality rate associated with lung cancer for each NUTS II was calculated based on INE data. The mortality rate was available for 2013 and 2014, the years of diagnosis considered, until 2019.

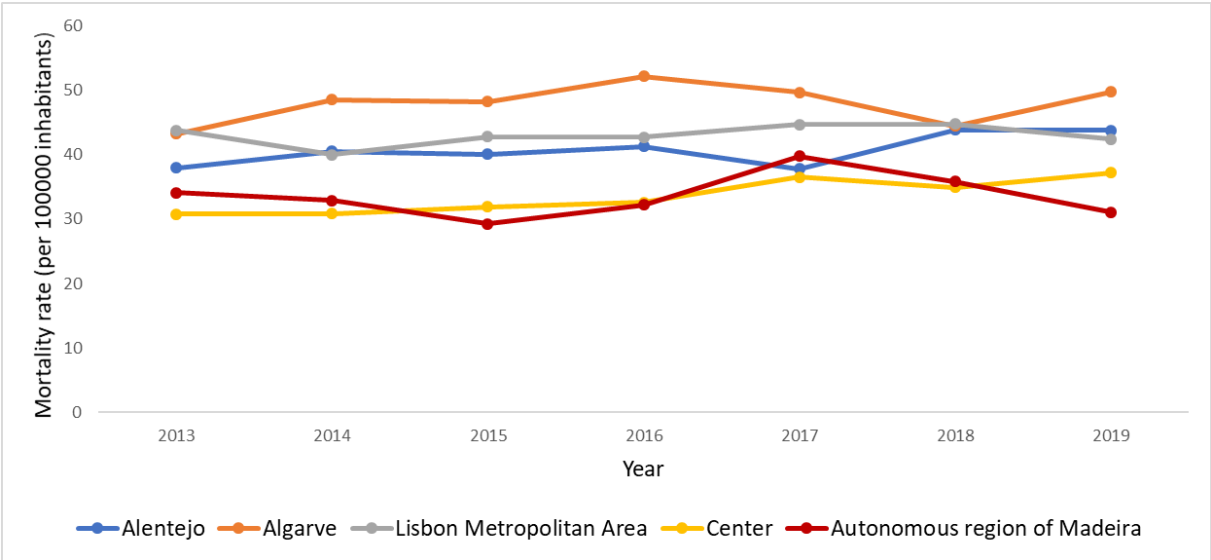

Figure 4.19: Mortality rate per 100000 inhabitants for each NUTS II region.

The mortality rate per 100000 inhabitants in Portugal had some variations in the different NUTS II regions over time. Algarve was the region where the mortality rate was higher since 2014 when compared

with the remaining NUTS II regions. Additionally, in 2018 the region with the highest mortality rate was Lisbon Metropolitan Area. In 2019, with 31 deaths per 100000 inhabitants, the Autonomous Region of Madeira was the region with the lowest mortality associated lung cancer. In the same year, with 50 deaths per 100000 inhabitants, the highest mortality rate was registered in Algarve.

Based on Figure 4.19, 2013 and 2018 have lower variability between the NUTS II regions. In the remaining years, there was greater variability in the mortality rate between these regions.

4.3 Ordinal Logistic Regression

This analysis was made considering all variables previously discussed, that is, stage as dependent variable and gender, age, doc1000, benef, att1000 and income.index as independent variables. Univariate regressions for each variable are also presented.

Unlike the usual models, this type of model was characterized by more than one equation. As seen in section 3.2, there were as many equations as there were categories of the dependent variable, minus one. To simplify the notation, the categories of the response variable I, II, IIIA and IIIB were represented in the scale 1, 2, 3 and 4, respectively. So, in this case, in order to evaluate the model the following four equations were interpreted:

Equation 1: $logit(P(Y_i \leq 1)) = \alpha_1 - \beta\mathbf{X} \longrightarrow$ comparison I *vs.* II, IIIA, IIIB, IV

Equation 2: $logit(P(Y_i \leq 2)) = \alpha_2 - \beta\mathbf{X} \longrightarrow$ comparison I, II *vs.* IIIA, IIIB, IV

Equation 3: $logit(P(Y_i \leq 3)) = \alpha_3 - \beta\mathbf{X} \longrightarrow$ comparison I, II, IIIA *vs.* IIIB, IV

Equation 4: $logit(P(Y_i \leq 4)) = \alpha_4 - \beta\mathbf{X} \longrightarrow$ comparison I, II, IIIA, IIIB *vs.* IV

Firstly, the simplest models were fitted, in which only one independent variable was used to model the dependent variable. The purpose of a simple model was to provide a simple low-dimensional summary of a dataset, in which the individual relation between independent and dependent variables was assessed. As seen in the chapter 2, the variables used in statistical inference were age, gender, doc1000, att1000, benef and income.index. Only age and gender corresponded to unique characteristics that define the individual. A little more individualized, but still built based on information whose aggregating element was also the residence county, there was income.index. The remaining variables were doc1000, att1000 and benef and these did not characterize the patient individually, since those who lived in the same county had the same value in these indicators. Thus, the simplest model applied to individual information was cumulative link model whereas the area-based information, including the income.index, was modeled by cumulative link mixed model, where the random effect was added. This random effect represented the residence county and this was the major difference between these two models.

The main results are displayed Table 4.3. The cumulative link model expression is given by:

$$ln\left(\frac{P(Y_i \leq j)}{1 - P(Y_i \leq j)}\right) = ln\left(\frac{P(Y_i \leq j)}{P(Y_i > j)}\right) = \alpha_j - \beta \times x_i \tag{4.1}$$

where $i = 1, ..., 2266$, $j = 1, 2, 3, 4$ and x corresponds to the set of individual independent variables used to explain the severity of NSCLC. The cumulative link mixed model expression is given by:

$$ln\left(\frac{P(Y_i \leq j)}{1 - P(Y_i \leq j)}\right) = ln\left(\frac{P(Y_i \leq j)}{P(Y_i > j)}\right) = \alpha_j - \beta \times x_i - u(county_i) \tag{4.2}$$

where $i = 1, ..., 2266$, $j = 1, 2, 3, 4$ and x corresponds to the set of area-based indicators mentioned above.

Table 4.3: Summarized information of each univariate model.

	$\hat{\alpha}_1$	$\hat{\alpha}_2$	$\hat{\alpha}_3$	$\hat{\alpha}_4$	$\hat{\beta}$	$\hat{\sigma}^2$	$p-value$	95% CI(β)
genderMale*	-1.985	-1.499	-0.853	-0.431	-0.076	-	0.427	(-0.266 , 0.111)
age*	-2.219	-1.732	-1.087	-0.665	-0.004	-	0.233	(-0.012 , 0.003)
doc1000**	-2.099	-1.610	-0.962	-0.538	-0.029	0.006	< 0.01	(-0.046 , -0.011)
benef**	-2.110	-1.621	-0.971	-0.545	-0.003	0.033	0.293	(-0.009 , 0.003)
att1000**	-1.744	-1.257	-0.611	-0.189	0.0001	1	0.286	(-0.00009 , 0.0002)
income.index**	-2.190	-1.701	-1.050	-0.625	-0.029	0.023	0.029	(-0.055 , -0.003)

* Cumulative link model
** Cumulative link mixed model

Since the regression coefficient is positive the possibility of the stage increasing to a higher stage increases with the increase of the independent variable. Otherwise, the possibility ofa diagnosis at higher stages decreases with the increase of the independent variable. Both 95% confidence interval and the $p-value$ were indicators to assess the significance of each independent variable.

Taking into consideration the sign of the regression coefficients ($\hat{\beta}$), considering one variable at a time, it was possible to observe that gender and age of the patient, the number of doctors, the welfare recipients and the estimated annual income favor lower stages. On the other hand, the number of attendances favor later stages. Based on $p-values$, the number of doctors and the estimated annual income are significant variables, without any effect caused by other independent variables.

The next step consisted in estimating a multivariable proportional odds model and the results are displayed in Table 4.4. In this first phase, the area-based information was treated as individual, thus making unnecessary the inclusion of a random effect. The main purpose of this modeling approach was to compare the results with the model in which the aggregator element as a random term was included.

Table 4.4: Summarized information of multivariable proportional odds model.

	$\hat{\alpha}_1$	$\hat{\alpha}_2$	$\hat{\alpha}_3$	$\hat{\alpha}_4$	$\hat{\beta}$	$p-value$	95% CI(β)
genderMale	-2.317	-1.828	-1.179	-0.755	-0.143	0.419	(-0.502 , 0.209)
age					-0.004	0.335	(-0.011 , 0.004)
doc1000					-0.029	0.089	(-0.063 , 0.004)
benef					-0.002	0.577	(-0.007 , 0.004)
att1000					0.0001	0.312	(-0.00005 , 0.0002)
income.index					0.008	0.773	(-0.048 , 0.064)

At this point, the proportional odds assumption should be evaluated. One of the assumptions underlying ordinal logistic regression was that the relationship between each pair of outcome groups would be the same. That is, the ordinal logistic regression assumes that the coefficients that describe the relationship between the lowest versus all higher categories of the response variable are the same as those that describe the relationship between the next lowest category and all higher categories, etc. Since the relationship between all pairs of groups was the same, only one set of coefficients is presented in Table 4.4. This assumption was tested applying a series of *Wald* tests to verify whether the regression coefficients differ across equations. With a $p-value$ lower than 0.01, gender was the only variable for

which this assumption was rejected. Thus, gender will have different effects in comparison between the stages at diagnosis.

Proceeding with the analysis, the partial proportional odds model with *logit* function was fitted with gender coefficients changing across the stages while for the other variables assume parallel lines assumption was kept. The general equation of the model is

$$\begin{aligned} logit(P(Y_i \leq j)) = \alpha_j - \beta_1(age_i) - \beta_2(doc1000_i) - \beta_3(benef_i) - \beta_4(att1000_i) - \\ \beta_5(income.index_i) + \tilde{\beta}_{6j}(gender_i), \quad i = 1,...,2266, j = 1,2,3,4 \end{aligned} \tag{4.3}$$

In this model, thirteen coefficients were to be estimated: four thresholds, four coefficients associated to the gender variable, which varied across the stage at diagnosis, and one for each of the remaining variables. The obtained results are as follows:

Table 4.5: Summarized information of multivariable partial proportional odds model.

	$\hat{\alpha}_1$	$\hat{\alpha}_2$	$\hat{\alpha}_3$	$\hat{\alpha}_4$	$\hat{\beta}$	$p-value$	95% CI(β)
genderMale	-1.925	-1.677	-1.184	-0.859	(1: -0.457, 2: -0.108, 3: 0.102, 4: 0.235)	(1: 0.027, 2: 0.580, 3: 0.582, 4: 0.195)	-
age					-0.004	0.315	(-0.011 , 0.004)
doc1000					-0.026	0.131	(-0.059 , 0.008)
benef					-0.002	0.519	(-0.007 , 0.004)
att1000					0.0001	0.306	(-0.00005 , 0.0002)
income.index					0.003	0.917	(-0.053 , 0.059)

The coefficients for age, doc1000 and benef are negative, while the coefficients for income.index and att1000 are positive. Taking into account the coefficient sign of each variable, the results indicate that:

- the odds of a patient being diagnosed at higher stages decreases with the increase of the patients' age, the number of doctors and the number of welfare recipients in the patients' county of residence
- the odds of a patient being diagnosed at higher stages increases with the increase of the patients' income and the number of attendances in the patients' county of residence

The interpretation for gender varies depending on the category of the response variable, stage at diagnosis. Gender was a categorical variable with two categories – *Female* and *Male* – where female was the reference class. For the lowest stages ($Y \leq 2$), the possibility of a male patient being diagnosed at a higher stage was higher than that of a female patient. For higher stages ($Y \geq 3$), the conclusion was the opposite, the odds of a female patient being diagnosed at a higher stage was higher than that of a male patient.

In order to understand the effect of gender in the model it could be interesting analyse the $P(Y > j)$ with $j = 1,2,3,4$ for each value that the variable can takes - *Female* or *Male*.

$$
\begin{aligned}
&\frac{P(Y \leq j|gender)}{P(Y > j|gender)} = \exp(\alpha_j + \beta_j \times gender) \\
&\Leftrightarrow P(Y \leq j|gender) = \exp(\alpha_j + \beta \times gender) P(Y > j|gender) \\
&\Leftrightarrow P(Y \leq j|gender) = \exp(\alpha_j + \beta_j \times gender)(1 - P(Y \leq j|gender)) \\
&\Leftrightarrow P(Y \leq j|gender) = \exp(\alpha_j + \beta_j \times gender) - \exp(\alpha_j + \beta_j \times gender)(P(Y \leq j|gender) \\
&\Leftrightarrow P(Y \leq j|gender)(1 + \exp(\alpha_j + \beta_j \times gender)) = \exp(\alpha_j + \beta_j \times gender) \\
&\Leftrightarrow P(Y \leq j|gender) = \frac{\exp(\alpha_j + \beta_j \times gender)}{1 + \exp(\alpha_j + \beta_j \times gender)}
\end{aligned}
\tag{4.4}
$$

Note that when j was the first level, $P(Y \leq j) = P(Y = j)$. However, to calculate $P(Y = j+1)$, the following expression was used:

$$
\begin{aligned}
&P(Y \leq j+1|gender) = P(Y = j|gender) + P(Y = j+1|gender) \\
&\Leftrightarrow P(Y = j+1|gender) = P(Y \leq j+1|gender) - P(Y = j|gender) \\
&\Leftrightarrow P(Y = j+1|gender) = \frac{\exp(\alpha_{j+1} + \beta_{j+1} \times gender)}{1 + \exp(\alpha_{j+1} + \beta_{j+1} \times gender)} - P(Y = j|gender)
\end{aligned}
\tag{4.5}
$$

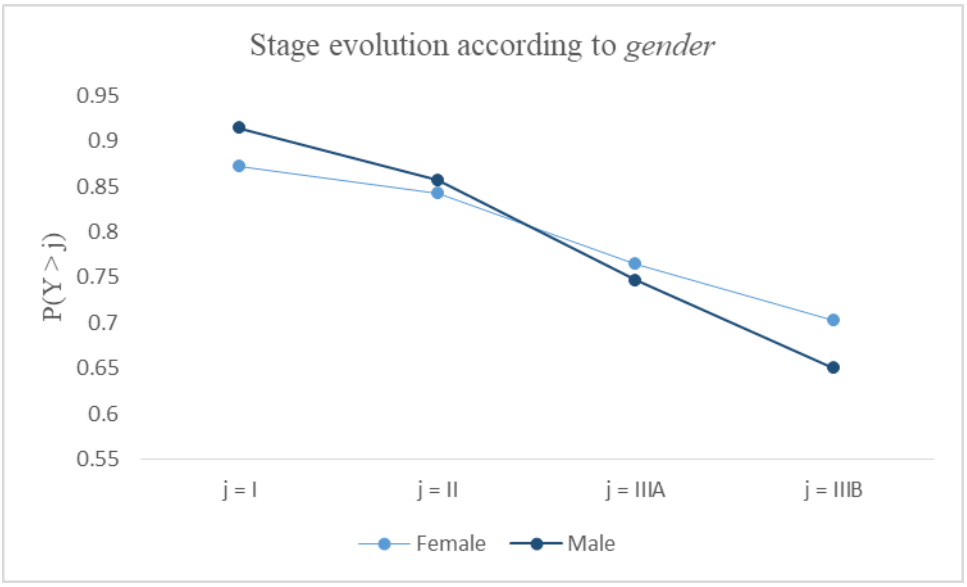

Figure 4.20: Effect of gender in the cumulative link model.

Figure 4.20 displays the probability of a patient being diagnosed at higher stages for each considered cut-point according to the gender:

$P(Y > 1) \Longrightarrow$ I *vs.* II, IIIA, IIIB, IV

$P(Y > 2) \Longrightarrow$ I, II *vs.* IIIA, IIIB, IV

$P(Y > 3) \Longrightarrow$ I, II, IIIA *vs.* IIIB, IV

$P(Y > 4) \Longrightarrow$ I, II, IIIA, IIIB *vs.* IV

As previously mentioned, for lower stages - I and II, the probability of a male patient being diangosed at higher stages is higher when compared with a female patient. However, the conclusion was opposite when the severity of disease increases.

Odds Ratio

Calculating the OR, as presented in section 3, for each variable, the results are the following:

Age: $OR = \frac{odds_{age+1}}{odds_{age}} = \exp(\hat{\beta}_1) = \exp(-0.004) = 0.996$

Number of doctors: $OR = \frac{odds_{doc_{1000}+1}}{odds_{doc_{1000}}} = \exp(\hat{\beta}_2) = \exp(-0.026) = 0.974$

Welfare of recipients: $OR = \frac{odds_{benef+1}}{odds_{benef}} = \exp(\hat{\beta}_3) = \exp(-0.002) = 0.998$

Number of attendances: $OR = \frac{odds_{att_{1000}+100}}{odds_{att_{1000}}} = \exp(100 \times \hat{\beta}_4) = \exp(100 \times 0.00005) = 1.005$

Income: $OR = \frac{odds_{income+1}}{odds_{income}} = exp(\hat{\beta}_5) = \exp(0.003) = 1.003$

Gender:

- Stage I: $OR = \frac{odds_{Male}}{odds_{Female}} = \exp(-\hat{\beta}_{61}) = \exp(0.45) = 1.568$
- Stage II: $OR = \frac{odds_{Male}}{odds_{Female}} = \exp(-\hat{\beta}_{62}) = \exp(0.11) = 1.116$
- Stage IIIA: $OR = \frac{odds_{Male}}{odds_{Female}} = \exp(-\hat{\beta}_{63}) = \exp(-0.10) = 0.905$
- Stage IIIB: $OR = \frac{odds_{Male}}{odds_{Female}} = \exp(-\hat{\beta}_{64}) = \exp(-0.24) = 0.787$

The above values correspond to the OR of the event $Y \geq j$, with j = I, II, IIIA, IIIIB, IV.

According to the result for age (OR 0.996, 95% CI: 0.989 - 1.004) the odds of an older patient being diagnosed at higher stages was 0.996 times the odds of an younger patient being diagnosed at higher stages.

Regarding the number of doctors per 1000 inhabitants (OR 0.974, 95% CI: 0.942 - 1.008), the result indicated that the odds of being diagnosed at higher stages was 2.6% lower for patients who live in a county where the number of doctors per 1000 inhabitants was one unit higher.

The value for the number of welfare recipients per 1000 inhabitants (OR 0.998, 95% CI: 0.993 - 1.004) indicated that the odds of being diagnosed at higher stages decreases by 0.2% for patients who live in a county where the number of welfare recipients per 1000 inhabitants was one unit higher.The odds of a patient to be diagnosed at higher stages, if lived in a county in which the number of attendances per 1000 inhabitants was higher, was 0.5% higher than a patient who lives in a county where the number of attendances per 1000 inhabitants was lower (OR 1.005, 95% CI: 0.995 - 1.016).

The odds of a patient being diagnosed at higher stages increased by 0.3% by each income one thousand euro increasing (OR 1.003, 95% CI: 0.949 - 1.060).

Since the proportional odds assumption was not verified for gender, its coefficients were different across the cut-points/thresholds. Hence, the OR were different too. For stage I the OR indicated that the

odds of being diagnosed at higher stages (II, IIIA, IIIB or IV) was 56.8% higher for male patients. For stage II, the odds of being diagnosed at higher stages (IIIA, IIIB or IV) for a male patient was 11.6% higher comparing with the odds of this event for a female patient. The result for stage IIIA indicated that the odds of a male patient being diagnosed at higher stages (IIIB or IV) was 9.5% lower than a female patient. Considering the stage IIIB or below, the odds of a male patient being diagnosed at higher stages (IV) was 21.3% lower when compared to a female patient.

Excluding gender, the OR were very close to one, meaning that there is no evidence that exposure (independent variable) significantly affects the odds of outcome (response variable).

Cumulative link mixed models

In equation (4.2), **x** corresponds to the set of variables used as independent variables. As in the equation (4.1), age, doc1000, benef, att1000 and income.index are the variables used for univariate analysis.

As in the previous simplest models (see Table 4.3), the number of attendances was the only variable whose regression coefficient was positive. There was a slight difference about the significance of each these variables. The $p-value$ associated to the number of attendances and income increases when the random effect was added. The complete model was fitted and its expression is given by:

$$\begin{aligned} logit(P(Y_i \leq j)) = \alpha_j - \beta_1(age_i) - \beta_2(doc1000_i) - \beta_3(benef_i) - \beta_4(att1000_i) \\ - \beta_5(income.index_i) - \beta_6(gender_i) - u(county_i) \\ i = 1, ..., 2266; j = 1, 2, 3, 4 \end{aligned} \tag{4.6}$$

The *R* function nominal_test() used to test available to test the proportional odds assumption is not available for mixed models. Hence, the mixed model (4.6) and the mixed model with a nominal effect were compared. To compare two nested models, or a sequence of more than two nested models, the *R* function anova() was used. This comparison was made considering model (4.6) and the models with each of the independent variables varying according to the stage of diagnosis, for each variable separately. Hence, six comparisons were made, as many as existing independent variables. The *likelihood ratio* was computed, allowing to choose the best model between each combination of two nested models. The results are displayed in Table 4.6.

Table 4.6: Analysis of the proportional odds assumption for the cumulative link mixed model.

Variable whose proportional odds assumption is tested	LR	*p-value*
age	0.899	0.343
doc1000	1.796	0.180
benef	-0.309	1
att1000	-71.601	1
income.index	-0.003	1
gender	0.293	0.588

Table 4.6 presents the *p-values* for each variable of the test for the addition of a nominal effect. Larger *p-values* indicates that the difference between the two models is not statistically significant, and the simplest model was chosen. Smaller *p-values*, indicates that the difference was statistically significant, and the model with nominal effect was chosen. There were no *p-values* lower than 0.1, maximum

significance level used. Hence, the proportional odds assumption was verified for all the independent variables.

Table 4.7 displays the summarized information about the multivariable mixed model.

Table 4.7: Summarized information of each multivariable model including the effect of random term.

	$\hat{\alpha}_1$	$\hat{\alpha}_2$	$\hat{\alpha}_3$	$\hat{\alpha}_4$	$\hat{\beta}$	$\hat{\sigma}^2$	$p-value$	95% CI($\hat{\beta}$)
genderMale	-2.317	-1.828	-1.179	-0.755	-0.143	1	0.499	(-0.559 , 0.272)
age					-0.004		0.360	(-0.012 , 0.004)
doc1000					-0.029		0.566	(-0.128 , 0.070)
benef					-0.002		0.777	(-0.012 , 0.009)
att1000					0.00007		0.487	(-0.0001 , 0.0003)
income.index					0.008		0.825	(-0.06 , 0.075)

The interpretation of the mixed model was made based on the sign of coefficients, in which the larger coefficients reveals an association with larger scores.

- the odds of an older patient to be diagnosed at higher stages was lower than a younger patient
- higher income was associated to the increase of the odds of being diagnosed at higher stages
- patients who live in counties with a high number of welfare recipients had a lower odds of being diagnosed at higher stages than patients who live in counties with a smaller number of welfare recipients
- the odds of a patient to be diagnosed at higher stages was higher if his residence county had a higher number of attendances
- the odds of a male patient to be diagnosed at higher stages was lower than a female patient
- patients who live in counties with a high number of doctors had a lower odds of being diagnosed at higher stages than patients who live in counties with a smaller number of doctors

Unlike the model without random effect, the model with random effects had no variables whose proportional odds assumption was not verified, including gender. This variable did not have the same effect at different stages as the Figure 4.20 suggested. Adding the county as random effect, the effect of gender between the stages was different. In Figure 4.21 one can observe that the distance between the lines is shorter and, more importantly, do not intercept, compared to the Figure 4.20, suggesting parallelism between the two lines. Indeed, as seen in chapter 3, the proportional odds assumption was also known as parallel lines assumption.

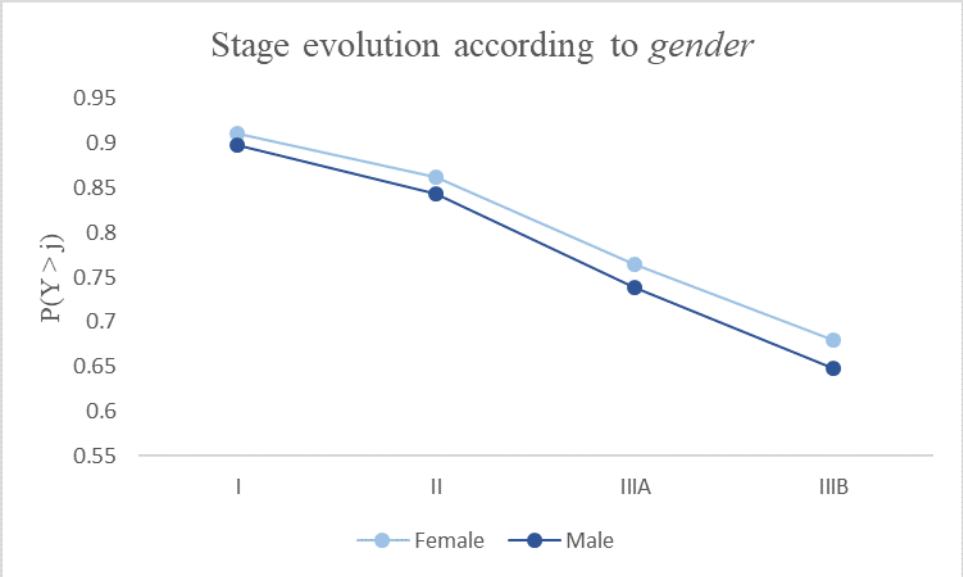

Figure 4.21: Effect of gender in the cumulative link mixed model.

The dark blue line indicates the probability of a male patient to be diagnosed at higher stages is lower than lower stages, as shown in Table 4.7.

Odds Ratio

Age: $OR = \frac{odds_{age+1}}{odds_{age}} = exp(\hat{\beta}_1) = exp(-0.004) = 0.996$

Number of doctors: $OR = \frac{odds_{doc_{1000}+1}}{odds_{doc_{1000}}} = exp(\hat{\beta}_2) = exp(-0.029) = 0.971$

Welfare of recipients: $OR = \frac{odds_{benef+1}}{odds_{benef}} = exp(\hat{\beta}_3) = exp(-0.002) = 0.998$

Number of attendances: $OR = \frac{odds_{att_{1000}+100}}{odds_{att_{1000}}} = exp(100 \times \hat{\beta}_4) = exp(100 \times 0.00007) = 1.007$

Income: $OR = \frac{odds_{income+1}}{odds_{income}} = exp(\hat{\beta}_5) = exp(0.008) = 1.008$

Gender: $OR = \frac{odds_{Male}}{odds_{Female}} = exp(\hat{\beta}_6) = exp(-0.143) = 0.866$

The OR associated with the age indicated that the chance of being diagnosed in higher stages decreased 0.4% for older patients. The chance of a patient being diagnosed at higher stages decreased by 2.9% if his residence county had a larger number of doctors per 1000 inhabitants. Patients living in counties with a high number of welfare recipients had a 0.2% lower chance of being diagnosed at higher stahes than patients living in counties with a smaller number of welfare recipients. The OR for att1000 was greater than 1, indicating that the chance of a patient being diagnosed at higher stages was higher for those who live in counties with the number of attendances per 1000 inhabitants higher. The estimated annual income had also a OR greater than one, indicating that patients with higher estimated annual income had a 0.8% increase in the chance of being diagnosed in higher stages. Finally, the OR for gender

was 0.866, indicating that the chance of a male patient being diagnosed at higher stages decreased 13.4% when compared to a female patient.

The inclusion of county as a random effect allowed to account for the variability that the independent variables cannot explain, to incorporate county-to-county variability of NSCLC, and improve the ability to describe how fixed effects relate to outcomes.

The estimated standard deviation ($\hat{\sigma}_u = 1$) of the random effect indicates the existence of variability caused by the residence of the patient.

The conditional modes and the conditional variance, which provides an uncertainty measure of the conditional modes, were used to assess the effect of random term. The conditional modes are an available component of the cumulative link mixed model (ranef()). These values correspond to the difference between the average predicted response for a given set of fixed-effect values and the response predicted for a particular individual.

Figure 4.22 characterizes the county as a random effect through the conditional modes as well their 95% confidence intervals. Figure 4.22 displays 116 lines, corresponding to each county that cover the ROR-Sul area. Each colour corresponds to the NUTS II region of the county.

Table 4.8: NUTS II cataloged by colored considered.

Center	red
Alentejo	blue
Lisbon Metropolitan Area	orange
Algarve	yellow
Autonomous Region of Madeira	green

Table 4.9: Correspondence of indexes to the respective counties.

1	Abrantes	30	Calheta (R.A.M.)	59	Mação	88	Ribeira Brava
2	Alandroal	31	Câmara De Lobos	60	Machico	89	Rio Maior
3	Albufeira	32	Campo Maior	61	Mafra	90	Salvaterra De Magos
4	Alcácer Do Sal	33	Cartaxo	62	Marvão	91	Santa Cruz
5	Alcanena	34	Cascais	63	Mértola	92	Santana
6	Alcobaça	35	Castro Marim	64	Moita	93	Santarém
7	Alcochete	36	Castro Verde	65	Monchique	94	Santiago Do Cacém
8	Alenquer	37	Chamusca	66	Montemor-O-Novo	95	São Brás De Alportel
9	Aljezur	38	Constância	67	Montijo	96	Sardoal
10	Aljustrel	39	Coruche	68	Mora	97	Seixal
11	Almada	40	Crato	69	Moura	98	Serpa
12	Almeirim	41	Elvas	70	Mourão	99	Sesimbra
13	Almodôvar	42	Entroncamento	71	Nazaré	100	Setúbal
14	Alpiarça	43	Estremoz	72	Nisa	101	Silves
15	Alter Do Chão	44	Évora	73	Óbidos	102	Sines
16	Alvito	45	Faro	74	Odemira	103	Sintra
17	Amadora	46	Ferreira Do Alentejo	75	Odivelas	104	Sobral De Monte Agraço
18	Arraiolos	47	Ferreira Do Zêzere	76	Oeiras	105	Tavira
19	Arronches	48	Fronteira	77	Olhão	106	Tomar
20	Arruda Dos Vinhos	49	Funchal	78	Ourém	107	Torres Novas
21	Avis	50	Gavião	79	Ourique	108	Torres Vedras
22	Azambuja	51	Golegã	80	Palmela	109	Vendas Novas
23	Barreiro	52	Grândola	81	Peniche	110	Viana Do Alentejo
24	Beja	53	Lagoa	82	Ponte De Sôr	111	Vidigueira
25	Benavente	54	Lagos	83	Portalegre	112	Vila Do Bispo
26	Bombarral	55	Lisboa	84	Portel	113	Vila Franca De Xira
27	Borba	56	Loulé	85	Portimão	114	Vila Nova Da Barquinha
28	Cadaval	57	Loures	86	Redondo	115	Vila Real De Santo António
29	Caldas Da Rainha	58	Lourinhã	87	Reguengos De Monsaraz	116	Vila Viçosa

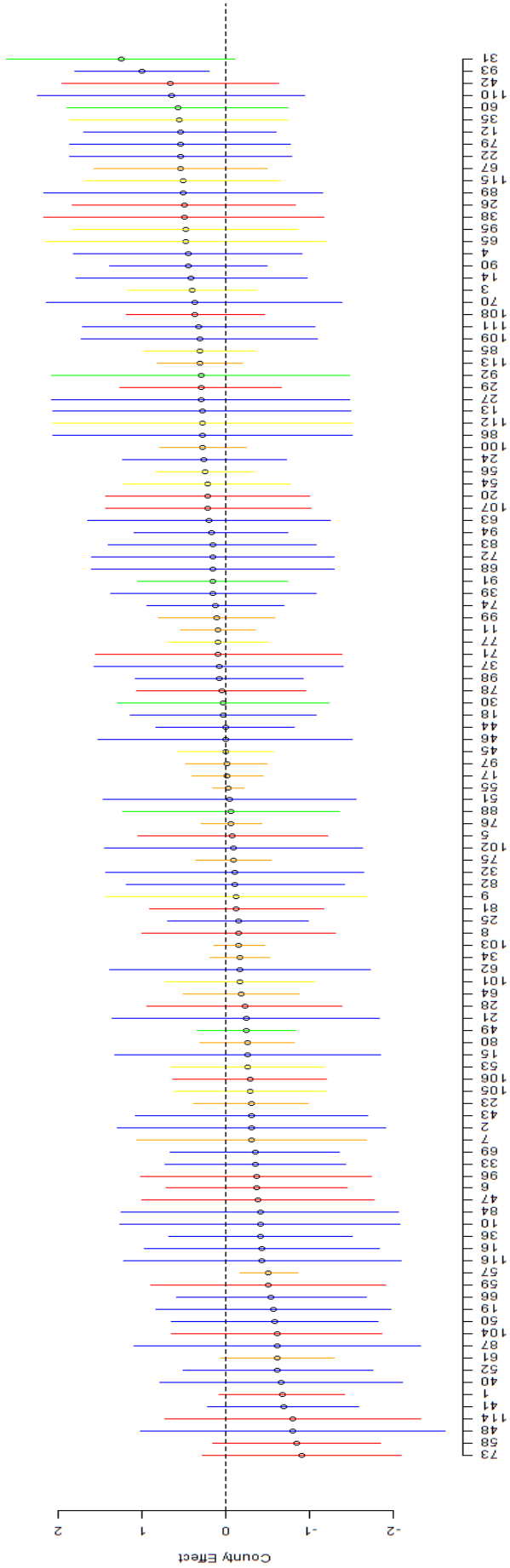

Figure 4.22: County effects given by conditional modes with 95% confidence intervals based on the conditional variance.

Figure 4.22 displays the conditional modes in ascending order. According to the results, in Óbidos (index 73) smaller stages were favoured. One the other hand, in Câmara dos Lobos (index 31) higher stages were favoured. The significant county effect indicates a difference in the incidence of NSCLC between the counties. Loures and Santarém (indexes 57 and 93, respectively) were the only counties whose 95% confidence intervals do not contain zero.

Additionally, based on Figure 4.22, most counties have a high amplitude and, therefore, the results should be analyzed carefully. The existence of high deviations in several counties was due to the fact that there were few cases of cancer diagnosed in those same counties. A small size contributes to an increase in deviation and the Figures 4.23 and 4.24 demonstrates the deviation. Figure 4.23 contains the total patients diagnosed with NSCLC by county. Lisbon was the county with more patients being diagnosed with NSCLC and Sintra was the second one. Indeed, Lisbon and the neighboring counties had the highest values for diagnosed cases. These counties also have smaller deviation.

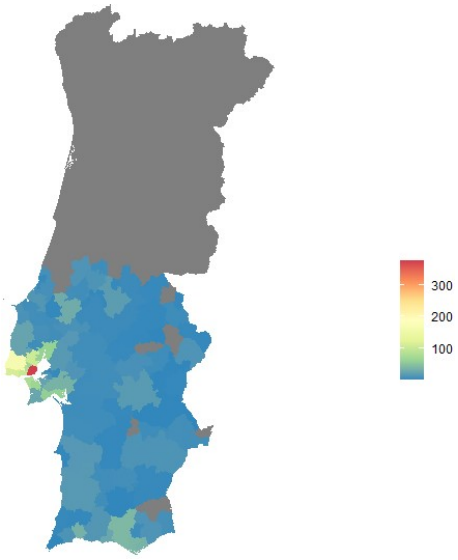

Figure 4.23: Number of diagnosed patients by county of ROR-Sul.

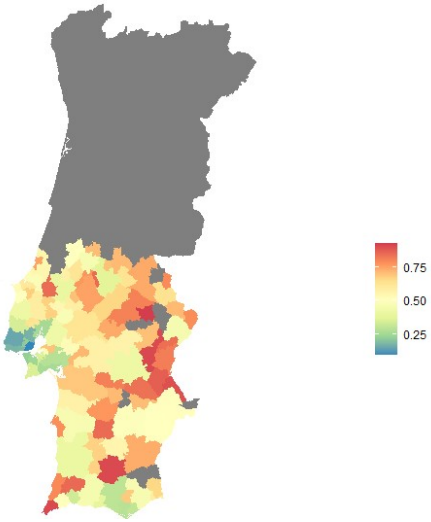

Figure 4.24: Square root of conditional variance of random effect.

This analysis was made using the residence county because both indicators of socioeconomic conditions and access to healthcare conditions were collected at the county level.

Even with the random effect, and although there were differences in incidence by residence county, the socioeconomic and access to healthcare conditions used in this project did not have a significant impact in stage at diagnosis of NSCLC.

Chapter 5

Conclusion and Discussion

The main purpose of this project was to study the association between the indicators of socioeconomic conditions and access to healthcare and the severity of lung cancer in patients diagnosed in the Southern region of Portugal.

The original dataset had demographic information about each patient, such as gender, age and place of residence (county and district) and specific information about the disease, such as the stage at diagnosis and the final status (alive or deceased).

The dependent variable considered was stage at diagnosis, which indicates the severity of the disease. Stage at diagnosis is an ordinal variable with more than two categories, hence, two different models were applied - the ordinal regression model and the ordinal regression mixed model. The main difference between these two models is the inclusion of a random effect. The random effect allows to accommodate the differences between the regions and, also, to capture what the variables cannot capture by themselves. The random term used in this model was the residence county, since this indicator was the aggregating element used to merge the socioeconomic and the access healthcare conditions to the original data.

From 01-01-2013 to 31-12-2014, 2266 patients living in the ROR-Sul area were diagnosed with NSCLC. Most of these patients were male (74.2%). The distribution of the proportion of individuals diagnosed in each stage showed differences depending on the district of residence. Regarding final status, 82.1% of patients died and, although the cause of death is not known, the proportion of patients diagnosed at stage IV who died is about five times higher than the proportion of patients diagnosed at stage IV who survived.

Socioeconomic and healthcare access information was not available at individual level and hence was gathered at the county level, based on official data sources. For that reason, patients living in the same county had the same value in these indicators, except for the income.index that was also added according to age and gender. These aspects of the data also justify the adoption of a spatial clustering component.

The impact of the chosen explanatory variables - gender, age, doc1000, att1000, benef and income.index - in stage at diagnosis was assessed through the cumulative link model. However, since variable gender has different effects across the cut-points (the proportional odds assumption was violated), the partial proportional odds model was applied (4.3). In these models, the difference between regions was not taken into consideration. To accommodate the dependency within regions, the county was introduced in the model as a random effect (4.6). None of the considered variables were found to have a statistically significant effect. When fitted the mixed effects model, the coefficients signs did not changes, indicating that the effect of the variables was the same regardless of the random term. The odds of a patient being diagnosed at higher stages decreased as age increased. Also, higher number of doctors per 1000 inhabitants and higher number of welfare recipients per 1000 inhabitants appeared as

favouring lower stages. Conversely, a higher number of attendances per 1000 inhabitants and a higher annual income appeared as contributing to an increased odds of diagnosis at a higher stage. In the model with no random term, the effect of patients' gender varies according to the severity of the disease. In the mixed model, the coefficient was negative, indicating that the odds of a male patient being diagnosed at later stages was lower when compared to a female patient. Unfortunately, none of the above mentioned effects was found to be statistically significant in either the models. However, when analysed in univariate models, both annual income and the number of doctors per 1000 inhabitants were found to be statistically associated to the severity at diagnosis. When considered in the multiple partial proportional odds model, despite loosing significance, variable doc1000 presented exactly the same coefficient estimate as in the univariate model.

Concerning variable income, in the univariate model, it appeared as being significantly associated to the stage at diagnosis, with higher income associated to less severe stages at diagnosis. When incorporated in the multiple models, not only it lost significance, but also the sign of the coefficient changed.

The lack of significance and also some contradiction that are found in our results may partially be explained by the type and quality of our data in terms of ability to properly characterize the patients and their socioeconomic conditions. According to several studies there is an association between access to healthcare and the socioeconomic status, which is characterized by social environment, lifestyle, occupation, education and income, among other factors [3, 11, 14, 15, 17, 35]. More precisely, potential factors as education, income, health insurance coverage, marital status have been studied to assess the association between indicators of socioeconomic and access to healthcare and lung cancer survival. Unfortunately, in this study, there was not much information available about access to healthcare, even in an aggregate format. The ideal would be to know each patient's situation, like, for example, if he or she has health insurance.

Considering the individualized or area-based studies, the results revealed that the lung cancer survival was lower for patients whose income is lower. Income is considered an important variable in health studies because, in general, a higher income is associated to a better access to healthcare and, consequently, better health outcomes. When such information is not available, education is often used since it is seen as a good proxy for income [37, 39]. In the present case, since none of the variables was available at the individual level, it was decided to extract information related to income from the official data sources and refine it at the finest categorization possible, accounting for location, age and gender. This was resumed in variables mabs and dgi, which were used to create variable income.index.

Gender has appeared in the literature as a risk factor, in which the lung cancer survival was lower for male patients. However, there is some evidence in the literature that suggests that for women, the risk of all major histological types of lung cancer is getting higher than for men. This difference could be justified with smoking habits. The prevalence of smoking has been increasing among women, while there is the opposite trend among men. Indeed, the results of the estimated mixed model (4.6) in this project revealed that the possibility of a male patient being diagnosed at higher stages was lower than to a female patient. As for the impact of smoking habits in the occurrence of the disease, there is a European study which confirms that "smoking has the same impact on lung cancer in the two sexes" [18].

In review articles, lung cancer survival has been the outcome of major interest and association was found with several socioeconomic conditions. Patients diagnosed at a higher stage have poorer prognosis which, in general, mean lower survival rate. Hence it is of major interest to identify the factors associated to the stage of the disease at the time of diagnosis, so that, eventually, some measures can be implemented with the objective of detecting the disease at earlier stages.

In the literature it was recommended the analysis combining the individual with aggregated data [11].

On the one hand, there are factors that belong exclusively to the individual, such as his or her income, age, gender, marital status, etc. On the other hand, access to healthcare factors already concerns the environment in which the patient is inserted, such as, number of available doctors. Most data used in this project correspond to the aggregated data.

Future research should consider patient-level characteristics such as smoking behaviours, comorbidities, socioeconomic status, education and marital status. Considering that lungs are the affected tissue, smoking habits might lead to higher vulnerability, and a lower chance survival [35].

Since the dependent variable is ordinal, there are several alternatives that could be used. The *continuation-ratio* model or *adjacent-category logistic* model are two examples of possible models that could be applied [6].

The *continuation-ratio* model was proposed by Feinberg [10] as an alternative method to the proportional odds model for the analysis of categorical data with ordered responses. The difference compared to the proportional odds model is that the cumulative probabilities, $P(Y \leq j)$, of being in one of the first j categories is replaced by the probability of being in category j, $P(Y = j)$.

The *adjacent-category logistic* model involves modelling the ratio of the two probabilities, $P(Y = j)$ and $P(Y = j+1)$ [5]. Exponentiating, the regression coefficient β_l, for the $l-th$ covariate x_l will result in the OR comparing $(Y = j)$ versus $(Y = j+1)$, for a unit increase in x_l.

Bibliography

[1] Carência socio-económica. rendimento social de inserção. Avalilable in: http://www.seg-social.pt/rendimento-social-de-insercao.

[2] Lung cancer: Symptoms, causes, diagnosis & treatment. Avalilable in: https://www.howtorelief.com/lung-cancer-symptom-causes-treatment/.

[3] Nancy E Adler and Katherine Newman. Socioeconomic disparities in health: pathways and policies. *Health affairs*, 21(2):60–76, 2002.

[4] Alan Agresti. *Categorical data analysis*, volume 482. John Wiley & Sons, 2003.

[5] Alan Agresti. *Analysis of ordinal categorical data*, volume 656. John Wiley & Sons, 2010.

[6] Cande V Ananth and David G Kleinbaum. Regression models for ordinal responses: a review of methods and applications. *International journal of epidemiology*, 26(6):1323–1333, 1997.

[7] Eric Bender. Epidemiology: The dominant malignancy. *Nature*, 513(7517):S2–S3, 2014.

[8] Lauren G Collins, Christopher Haines, Robert Perkel, and Robert E Enck. Lung cancer: diagnosis and management. *American family physician*, 75(1):56–63, 2007.

[9] Luc Djoussé, Joanne F Dorgan, Yuqing Zhang, Arthur Schatzkin, Maggie Hood, Ralph B D'Agostino, Donna L Copenhafer, Bernard E Kreger, and R Curtis Ellison. Alcohol consumption and risk of lung cancer: the framingham study. *Journal of the National Cancer Institute*, 94(24):1877–1882, 2002.

[10] Stephen E Fienberg. *The analysis of cross-classified categorical data*. Springer Science & Business Media, 2007.

[11] Isabelle Finke, Gundula Behrens, Linda Weisser, Hermann Brenner, and Lina Jansen. Socioeconomic differences and lung cancer survival—systematic review and meta-analysis. *Frontiers in oncology*, 8:536, 2018.

[12] James Gasperino. Gender is a risk factor for lung cancer. *Medical hypotheses*, 76(3):328–331, 2011.

[13] David W Hosmer Jr, Stanley Lemeshow, and Rodney X Sturdivant. *Applied logistic regression*, volume 398. John Wiley & Sons, 2013.

[14] Jan Hovanec, Jack Siemiatycki, David I Conway, Ann Olsson, Isabelle Stücker, Florence Guida, Karl-Heinz Jöckel, Hermann Pohlabeln, Wolfgang Ahrens, Irene Brüske, et al. Lung cancer and socioeconomic status in a pooled analysis of case-control studies. *PloS one*, 13(2), 2018.

[15] Aminah Jatoi, Paul Novotny, Stephen Cassivi, Matthew M Clark, David Midthun, Christi A Patten, Jeff Sloan, and Ping Yang. Does marital status impact survival and quality of life in patients with non-small cell lung cancer? observations from the mayo clinic lung cancer cohort. *The oncologist*, 12(12):1456–1463, 2007.

[16] Jolanta Lissowska, Alicja Bardin-Mikolajczak, Tony Fletcher, David Zaridze, Neonila Szeszenia-Dabrowska, Peter Rudnai, Eleonora Fabianova, Adrian Cassidy, Dana Mates, Ivana Holcatova, et al. Lung cancer and indoor pollution from heating and cooking with solid fuels: the iarc international multicentre case-control study in eastern/central europe and the united kingdom. *American journal of epidemiology*, 162(4):326–333, 2005.

[17] Jyoti Malhotra, Matteo Malvezzi, Eva Negri, Carlo La Vecchia, and Paolo Boffetta. Risk factors for lung cancer worldwide. *European Respiratory Journal*, 48(3):889–902, 2016.

[18] Moore Malvezzi, G Carioli, P Bertuccio, P Boffetta, F Levi, C La Vecchia, and E Negri. European cancer mortality predictions for the year 2017, with focus on lung cancer. *Annals of Oncology*, 28(5):1117–1123, 2017.

[19] Peter McCullagh. Regression models for ordinal data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 42(2):109–127, 1980.

[20] Nuno Miranda, Cristina Portugal, Paulo Jorge Nogueira, Carla Sofia Farinha, Ana Lisette Oliveira, Maria Isabel Alves, and José Martins. Portugal doenças oncológicas em números, 2015. *Portugal Doenças Oncológicas em números, 2015*, pages 7–65, 2016.

[21] IARC Working Group on the Evaluation of Carcinogenic Risks to Humans et al. Iarc monographs on the evaluation of carcinogenic risks to humans. ingested nitrate and nitrite, and cyanobacterial peptide toxins. *IARC monographs on the evaluation of carcinogenic risks to humans*, 94:v, 2010.

[22] World Health Organization. Cancer. Avalilable in: https://www.who.int/news-room/fact-sheets/detail/cancer.

[23] World Health Organization. Europe. Avalilable in: https://gco.iarc.fr/today/data/factsheets/populations/908-europe-fact-sheets.pdf, 2018.

[24] World Health Organization. Portugal. Avalilable in: https://gco.iarc.fr/today/data/factsheets/populations/620-portugal-fact-sheets.pdf, 2018.

[25] World Health Organization. World. Avalilable in: https://gco.iarc.fr/today/data/factsheets/populations/900-world-fact-sheets.pdf, 2018.

[26] World Health Organization et al. Gender in lung cancer and smoking research. 2004.

[27] Bercedis Peterson and Frank E Harrell Jr. Partial proportional odds models for ordinal response variables. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 39(2):205–217, 1990.

[28] Martin Pinquart and Paul R Duberstein. Associations of social networks with cancer mortality: a meta-analysis. *Critical reviews in oncology/hematology*, 75(2):122–137, 2010.

[29] JH Randall. The analysis of sensory data by generalized linear model. *Biometrical journal*, 31(7):781–793, 1989.

[30] M Patricia Rivera, Frank Detterbeck, and Atul C Mehta. Diagnosis of lung cancer: the guidelines. *Chest*, 123(1):129S–136S, 2003.

[31] Christopher G Slatore, David H Au, and Michael K Gould. An official american thoracic society systematic review: insurance status and disparities in lung cancer practices and outcomes. *American journal of respiratory and critical care medicine*, 182(9):1195–1205, 2010.

[32] American Cancer Society. About lung cancer. Avalilable in: https://www.cancer.org/cancer/lung-cancer/about/what-is.html.

[33] American Cancer Society. Non-small cell lung cancer stages. Avalilable in: https://www.cancer.org/cancer/lung-cancer/detection-diagnosis-staging/staging-nsclc.html.

[34] American Cancer Society. Treatment choices for non-small cell lung cancer, by stage. Avalilable in: https://www.cancer.org/cancer/lung-cancer/treating-non-small-cell/by-stage.html.

[35] C Martin Tammemagi, Christine Neslund-Dudas, Michael Simoff, and Paul Kvale. Smoking and lung cancer survival: the role of comorbidity and treatment. *Chest*, 125(1):27–37, 2004.

[36] Gerhard Tutz and Wolfgang Hennevogl. Random effects in ordinal regression models. *Computational Statistics & Data Analysis*, 22(5):537–557, 1996.

[37] R Valletta. Higher education, wages, and polarization. *Higher Education*, 2, 2015.

[38] Strother H Walker and David B Duncan. Estimation of the probability of an event as a function of several independent variables. *Biometrika*, 54(1-2):167–179, 1967.

[39] Scott A Wolla and Jessica Sullivan. Education, income, and wealth. *Page One Economics®*, 2017.

Appendices

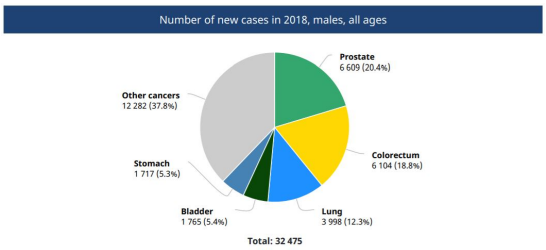

Figure 1: Portuguese number of new cases in 2018, male, all ages [24].

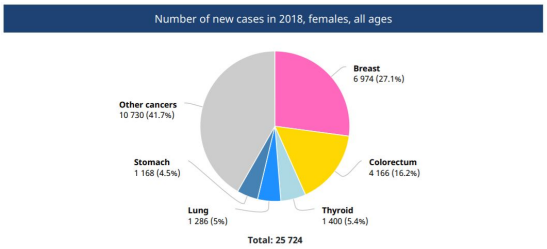

Figure 2: Portuguese number of new cases in 2018, female, all ages [24].

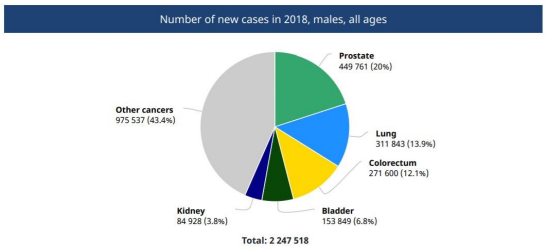

Figure 3: European number of new cases in 2018, male, all ages [23].

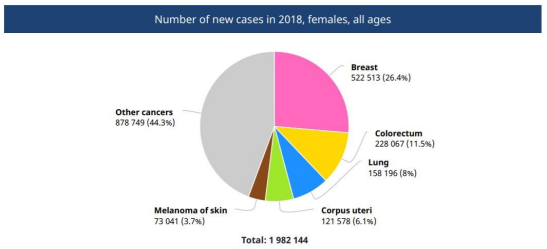

Figure 4: European number of new cases in 2018, female, all ages [23]

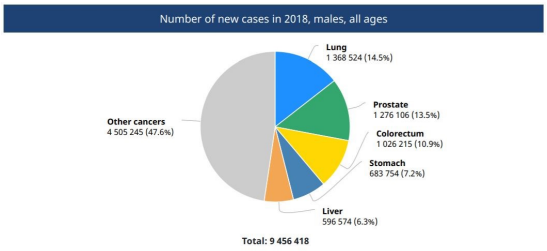

Figure 5: Worldwide number of new cases in 2018, male, all ages [25]

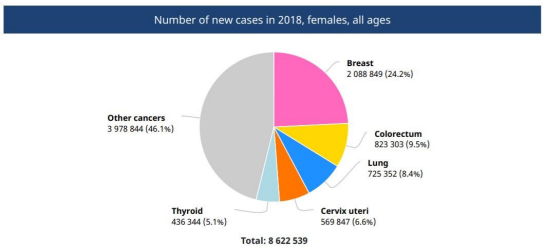

Figure 6: Worldwide number of new cases in 2018, female, all ages [25]

Proportional Odds Model

Example

A common example where this type of model is used is in wine data from Randall [29]. This dataset is available in the object wine in *R* package ordinal. It was used in a factorial experiment in order to determine the bitterness of wine where the lowest value (1) corresponds to "least bitter" classification and the highest value (5) corresponds to "most bitter". During wine production, two factors were used to evaluate the bitterness of wine - temperature and contact. Nine judges assessed wine from each of the total combinations (temperature - warm/could and contact - yes/no). The purpose of this project was to evaluate the effect of each factor on the perceived bitterness of wine.

The following cumulative link model for the wine data was considered:

$$logit(P(Y_i \leq j)) = \theta_j - \beta_1(temp_i) - \beta_2(contact_i) \quad i = 1,...,n \quad j = 1,...,J-1 \tag{1}$$

where i represents the number of observations and j the number of categories of response variable ($J = 5$). The model provides the cumulative probability of the $i-th$ observation falling in the $j-th$ category or

below.

The regression parameter β_1 corresponds to the impact in the bitterness of wine when the temperature is warm, keeping the contact unchanged. β_2 represents the impact in the bitterness of wine when the contact is "yes", keeping the temperature unchanged.

$$\hat{\theta_j} = \{-1.3444, 1.2508, 3.4669, 5.0064\}$$

$$\hat{\beta_1} = 2.5031 \quad \hat{\beta_2} = 1.5278$$

The regression coefficients are positive, indicating that increasing temperature and the existence of contact is associated with an increase in the bitterness of wine. $\hat{\theta_j}$ corresponds to the thresholds when $temp_i = cold$ and $contact_i = no$ are considered.

The three main conclusions of this model are as follows:

- The thresholds at $contact_i = yes$ conditions have been shifted a constant amount 1.5278 relative to the thresholds $contact_i = no$ at conditions.
- The location of the latent distribution has been shifted $+1.5278\sigma^*$ (scale units) at $contact_i = yes$ relative to $contact_i = no$.
- The OR of bitterness being rated in category j or above $(Y \geq j)$ is $exp(\hat{\beta_2}) = 4.61$

Partial or Non-Proportional Odds Model: nominal effects

Example

The cumulative link model in 1 specifies a structure in which the regression parameters are not allowed to vary with j. When the proportional odds assumption is relaxed in one variable, *contact* for example, the model expression is transformed by:

$$logit(P(Y_i \leq j)) = \theta_j - \beta(temp_i) + \tilde{\beta}_j(contact_i) \tag{2}$$

with $i = 1, ..., n$ and $j = 1, ..., J-1$. The obtained results with this model are as follows:

$$\hat{\theta_j} = \{-1.3230, 1.2464, 3.5500, 4.6602\}$$

$$\hat{\beta} = 2.519$$

$$\hat{\beta}_j = \{-1.6151, -1.5116, -1.6748, -1.0506\}$$

The thresholds vector refers to the $temp_i = cold$ and $contact_i = no$ settings while the thresholds at $temp_i = cold$ and $contact_i = no$ are $\hat{\theta}_j + \hat{\tilde{\beta}}_j$. Unlike the *proportional odds* model, the OR of bitterness being rated in category j or above now depends on j.

$$exp(-\hat{\tilde{\beta}}_j) = \{5.03, 4.53, 5.34, 2.86\}$$

To test the proportional odds assumption for all variables, the nominal_test() function is used. This function moves all terms in formula and copies all terms in scale to nominal one by one and produces a table with likelihood ratio tests of each term.

As seen previously, most of information is shared among patients and therefore this model can handle with some limitations. In this case, the assumption of independence of observations is not verified. This independence is assumed conditional on the fixed effect as well as the explanatory variable values.

Ordinal Generalized Linear Mixed Models

Example

Consider the same data of previously example [29]. The data used before as an example of cumulative link model was also analysed with mixed effects model by Tutz and Hennevogl [36].

The main objective was to find the factors that determines the bitterness of wine, taking into account the judge rating.The main purpose in using the judges as random effect is to deal with the given measures by each judge.

So, the cumulative link mixed model to the wine data is formulated as:

$$logit(P(Y_i \leq j)) = \alpha_j - \beta_1(temp_i) - \beta_2(contact_i) - u(judge_i) \tag{3}$$

with $i = 1, ..., n$ and $j = 1, ..., J-1$. The results contain the maximum likelihood estimates of the parameters:

$$\hat{\beta}_1 = 3.06, \hat{\beta}_2 = 1.83, \hat{\sigma}_u^2 = 1.29 = 1.13^2, \hat{\alpha_j} = \{-1.62, 1.51, 4.23, 6.09\}$$

The regression coefficients, β_1 and β_2, are positive indicating that with higher temperature and contact the bitterness of the wine increases. Analysing the temperature, the OR of the event $Y \geq j$ is $\exp(\beta(temp)) = \exp(3.06) = 21.37$.

The estimate of random effects corresponds to the standard deviation of this parameter which is given by $\sqrt{1.29} = 1.13$. The nullity test of this parameter does not appear in the output. However, it can be obtained through a likelihood ratio test. In order to assess the significance of this term a model that includes it and another model without it. The test of $\sigma_u = 0$ displays a p value showing whether the judge term is significant. Note that the $u_(judge_i)$ are not model parameters and they cannot be estimated in the conventional sense. Although, it is possible to assess its random effects. These ones are random and normally distributed $N(0, \sigma_u^2)$. In order to evaluate the impact of random effect there are two components of cumulative link mixed model that could be used - *ranef* and *condVar*. The first one corresponds to the random effect and the second one to the conditional variance. In this case, from these it is possible to know the judge effects via conditional modes with 95% confidence intervals.

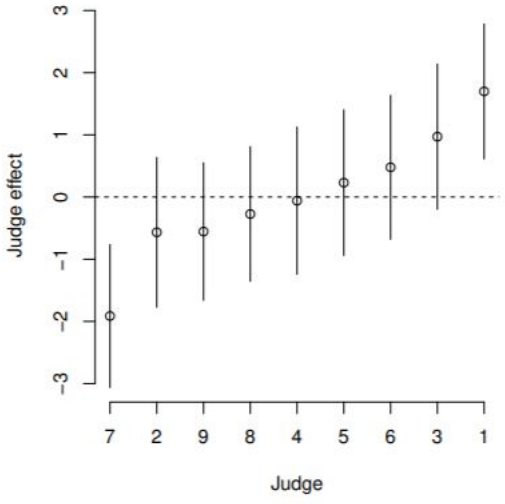

Figure 7: Judge effects given by conditional modes with 95% confidence intervals based on the conditional variance

Based on the figure 7 it is possible conclude the following:

- the lowest bitterness was given by seventh judge;
- the wine was classified as the most bitterness by first judge;
- based on the variation, the judges have different perceptions of bitterness.

R Code

```
> library(ordinal)
> mod_age <- clm(stage.cat ~ age, data = data)
> mod_doc1000 <- clm(stage.cat ~ doc1000, data = data)
> mod_att1000 <- clm(stage.cat ~ att1000, data = data)
> mod_benef <- clm(stage.cat ~ benef, data = data)
> mod_income <- clm(stage.cat ~ income.index, data = data)

# Proportional odds model

> library(ordinal)
> mod_pom <- clm(stage.cat ~ age + gender + doc1000 + benef + att1000 +
                   income.index, data=data)
> nominal_test(mod_pom) # to test the proportional odds assumption

# Output

Tests of nominal effects

formula: stage.cat ~ gender + age + doc1000 + benef + att1000 + income.index
             Df  logLik    AIC    LRT  Pr(>Chi)
<none>          -2775.6 5571.2
gender        3 -2759.5 5545.0 32.228 4.686e-07 ***
age           3 -2775.1 5576.3  0.937   0.81656
doc1000       3 -2774.2 5574.3  2.898   0.40754
benef         3 -2775.5 5576.9  0.290   0.96183
att1000       3 -2771.9 5569.8  7.364   0.06115 .
income.index  3 -2772.6 5571.2  6.031   0.11011
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1         1

# Partial proportional odds model

> mod_ppom <- clm(stage.cat ~ age + doc1000 + benef + att1000 + income.index,
           nominal=~gender, data=data)
> summary(mod_ppom)

formula: stage.cat ~ age + doc1000 + benef + att1000 + income.index
nominal: ~gender
data:    data

 link  threshold nobs logLik   AIC     niter max.grad cond.H
 logit flexible  2266 -2759.49 5544.98 6(2)  2.87e-09 1.4e+09

Coefficients:
```

```
               Estimate Std. Error z value Pr(>|z|)
age          -3.859e-03  3.852e-03  -1.002    0.316
doc1000      -2.559e-02  1.705e-02  -1.501    0.133
benef        -1.810e-03  2.788e-03  -0.649    0.516
att1000       5.274e-05  5.161e-05   1.022    0.307
income.index  2.496e-03  2.829e-02   0.088    0.930

Threshold coefficients:
                     Estimate Std. Error z value
I|II.(Intercept)       -1.9248     0.2985  -6.447
II|IIIA.(Intercept)    -1.6774     0.2955  -5.676
IIIA|IIIB.(Intercept)  -1.1836     0.2919  -4.055
IIIB|IV.(Intercept)    -0.8587     0.2908  -2.952
I|II.genderMale        -0.4565     0.2058  -2.218
II|IIIA.genderMale     -0.1081     0.1956  -0.553
IIIA|IIIB.genderMale    0.1021     0.1853   0.551
IIIB|IV.genderMale      0.2348     0.1813   1.295

# Proportional odds mixed model (county as random term)

> mod_pomm <- clmm(stage.cat ~ age + income.index + benef + att1000 + gender +
                     doc1000 +  (1|county), data = data)

> mod_pomm_age <- clmm(stage.cat ~ income.index + benef + att1000 + doc1000 +
                     gender +  (1|county), nominal = ~age, data = data)

> anova(mod_pomm, mod_pomm_age)

Likelihood ratio tests of cumulative link models:

             formula:
mod_pomm_age stage.cat ~ income.index + benef + att1000 + doc1000 + gender +
                (1 | county)
mod_ppom     stage.cat ~ age + income.index + benef + att1000 + gender + doc1000 +
                (1 | county)
             nominal: link: threshold:
mod_pomm_age ~age     logit flexible
mod_ppom     ~1       logit flexible

             no.par    AIC  logLik LR.stat df Pr(>Chisq)
mod_pomm_age     10 5644.6 -2812.3
mod_ppom         11 5645.7 -2811.9  0.8992  1      0.343

> mod_pomm_doc <- clmm(stage.cat ~ age + income.index + benef + att1000 +
                     gender + (1|county), nominal = ~doc1000, data = data)

> anova(mod_pomm, mod_pomm_doc)

Likelihood ratio tests of cumulative link models:

              formula:
mod_ppom_doc  stage.cat ~ age + income.index + benef + att1000 + gender +
            (1 | county)
mod_ppom      stage.cat ~ age + income.index + benef + att1000 + gender +
            doc1000 + (1 | county)
```

```
             nominal: link: threshold:
mod_ppom_doc ~doc1000 logit flexible
mod_ppom     ~1       logit flexible

             no.par    AIC  logLik LR.stat df Pr(>Chisq)
mod_ppom_doc     10 5645.5 -2812.8
mod_ppom         11 5645.7 -2811.9  1.7955  1     0.1803

> mod_pomm_att <- clmm(stage.cat ~ age + income.index + benef +  doc1000 +
                       gender + (1|county), nominal = ~att1000, data = data)

> anova(mod_pomm, mod_pomm_att)

Likelihood ratio tests of cumulative link models:

                   formula:
mod_pomm_att stage.cat ~ age + income.index + benef + doc1000 + gender +
                (1 | county)
mod_pomm     stage.cat ~ age + income.index + benef + att1000 + gender +
                doc1000 + (1 | county)
                   nominal: link: threshold:
mod_pomm_att ~att1000 logit flexible
mod_pomm     ~1       logit flexible

                   no.par    AIC  logLik LR.stat df Pr(>Chisq)
mod_pomm_att     10 5572.1 -2776.1
mod_pomm         11 5645.7 -2811.9 -71.601  1          1

> mod_pomm_income <- clmm(stage.cat ~ age + benef + att1000 + doc1000 +
                       gender +  (1|county), nominal = ~income.index, data = data)

> anova(mod_pomm, mod_pomm_income)

Likelihood ratio tests of cumulative link models:

                       formula:
mod_pomm_income stage.cat ~ age + benef + att1000 + doc1000 + gender +
                (1 | county)
mod_pomm        stage.cat ~ age + income.index + benef + att1000 + gender +
                doc1000 + (1 | county)
                       nominal:      link: threshold:
mod_pomm_income ~income.index logit flexible
mod_pomm        ~1            logit flexible

                       no.par    AIC  logLik LR.stat df Pr(>Chisq)
mod_pomm_income     10 5643.7 -2811.9
mod_pomm            11 5645.7 -2811.9 -0.0028  1          1

> summary(mod_pomm)

Cumulative Link Mixed Model fitted with the Laplace approximation

formula:
stage.cat ~ age + income.index + benef + att1000 + gender + doc1000 +
    (1 | county)
```

```
data:    data

 link  threshold nobs logLik   AIC     niter    max.grad cond.H
 logit flexible  2266 -2811.86 5645.73 298(923) 2.61e+02 5.9e+08

Random effects:
 Groups Name        Variance Std.Dev.
 county (Intercept) 1        1
Number of groups:  county 116

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
age          -3.701e-03  4.040e-03  -0.916    0.360
income.index  7.623e-03  3.451e-02   0.221    0.825
benef        -1.563e-03  5.528e-03  -0.283    0.777
att1000       6.843e-05  9.852e-05   0.695    0.487
genderMale   -1.433e-01  2.119e-01  -0.677    0.499
doc1000      -2.903e-02  5.062e-02  -0.573    0.566

Threshold coefficients:
          Estimate Std. Error z value
I|II       -2.3171     0.4541  -5.103
II|IIIA    -1.8279     0.4527  -4.038
IIIA|IIIB  -1.1791     0.4516  -2.611
IIIB|IV    -0.7547     0.4512  -1.673

# Maps in R
> library(rgdal)
> library(mptools)
> library(maps)
> library(mapproj)
> library(maptools)
> library(ggplot2)
> library(gpclib)
> library(tidyverse)

> gpclibPermit()

> myDF <- fortify(myFile, region = "NAME_2")

> mySHP <- choose.files()
> myFile <- readOGR(mySHP)
> myDF <- left_join(myDF, data, by='id')
> p <- ggplot(data = myDF, aes(x = long, y = lat, group = group, text =
    paste("Region:", county))) +
  geom_polygon(aes(fill = variable)) +
  scale_fill_distiller(palette = "Spectral") +
  theme(legend.position = "none") +
  theme(axis.title.x = element_blank()) +
  theme(axis.title.y = element_blank()) +
  theme_void() +
  labs(fill = "") +
  coord_map(xlim = c(-10, -5), ylim = c(36, 42.5))
```