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

The Predictive value of Microscopic Inflammation beyond The Endoscopic Margin at Diagnosis in Ulcerative Colitis Outcomes

Alexandra Sofia Ribeiro de Almeida

Mestrado em Bioestatística

Trabalho de Projeto orientado por:
Prof.ª Doutora Marília Cristina de Sousa Antunes
Doutora Joana Torres

2020

Resumo

A colite ulcerosa é uma doença inflamatória crónica do intestino, que se caracteriza por uma inflamação contínua da mucosa ("revestimento interno") do intestino grosso ou cólon. Esta doença é também caracterizada por períodos de remissão e exacerbações agudas, com um curso imprevisível da doença. Apesar da disponibilidade de novos medicamentos, ainda existe uma proporção considerável de pacientes que serão resistentes à terapia e por isso necessitam de colectomia ou desenvolvem complicações, destacando-se a necessidade de melhores marcadores preditivos do curso da doença.

Atualmente, a classificação da doença é baseada na classificação de Montreal, considerando a extensão endoscópica da doença no momento do diagnóstico, sendo os pacientes divididos em três subgrupos principais: inflamação limitada à zona do reto, ou proctite (E1), inflamação do lado esquerdo do intestino grosso (E2) e pancolite, que consiste na inflamação total do intestino grosso (E3). Este sistema de classificação tem valor prognóstico e implicações terapêuticas, dado que doentes com colites mais extensas têm normalmente quadros com curso clínico pior (maior necessidade de hospitalização, mais necessidade de cirurgia, maiores necessidades terapêuticas, etc). Contudo, a classificação de Montreal baseia-se apenas nos resultados endoscópicos ou no aspeto macroscópico no momento do diagnóstico. As recomendações atuais sugerem que sejam realizadas pelo menos duas biópsias em cinco segmentos diferentes do cólon, no momento do diagnóstico.

É frequente existir inflamação histológica ou microscópica acima da margem endoscópica, mas o seu significado prognóstico continua por descobrir. Contudo, estudos publicados sugerem que a presença de inflamação histológica nas colonoscopias de seguimento destes doentes é um importante factor prognóstico. A combinação de endoscopia e histologia parece fornecer uma indicação melhor da atividade da doença do que a endoscopia isolada, especialmente na mucosa endoscopicamente não inflamada. Nenhum estudo avaliou o valor prognóstico das características histológicas no momento do diagnóstico e, portanto, permanece desconhecido se os achados microscópicos pré-tratamento podem predizer o curso da doença.

Este tema de investigação foi proposto por duas médicas de gastrenterologia (Catarina Frias Gomes com orientação da doutora Joana Torres), que desenvolveram a parte clínica do projeto, bem como o protocolo de estudo. Esta dissertação, permitiu então responder às questões colocadas, desenvolvendo a componente estatística (com orientação da professora doutora Marília Antunes). Aqui, desenvolveu-se um estudo retrospectivo multicêntrico para avaliar o valor prognóstico das características histológicas em doentes com proctite e colite ulcerosa do lado esquerdo. Os objetivos deste estudo passam por avaliar a prevalência da inflamação microscópica na mucosa macroscópica não inflamada, bem como o valor preditivo das características histológicas da doença (na mucosa endoscopicamente inflamada e não inflamada) no momento do diagnóstico, a médio e longo prazo. Atendendo à necessidade cada vez mais importante de estratificar os doentes de acordo com os seus fatores de mau prognóstico, com vista a melhor seleção da terapêutica no momento do diagnóstico, ter um melhor conhecimento deste fator seria crucial. Neste estudo foram incluídos doentes com diagnóstico de proctite (E1) ou colite esquerda (E2)

entre 2006 e 2018 de diferentes centros hospitalares. Em Portugal tem havido um consumo considerável de recursos médicos associados à colite ulcerosa. Entre 2010 e 2017, cerca de 3400 pessoas foram hospitalizadas devido a esta doença, com uma estadia hospitalar média de 10 dias com um mínimo de 0 dias (para os indivíduos que são hospitalizados e não perfazem 24 horas) e um máximo de 109 dias. Assim, o estudo das hospitalizações devido a esta doença tornou-se uma questão muito relevante para se ter uma perspectiva da sua ocorrência, bem como para ajudar na monitorização dos pacientes e na tomada de decisões clínicas e de administração hospitalar. Neste trabalho, analisou-se o tempo até à hospitalização de 93 pacientes diagnosticados com colite ulcerosa nas fases E1 e E2.

Os modelos Cox foram utilizados uma vez que são versáteis, populares entre os investigadores médicos, especialmente no que diz respeito à interpretação dos parâmetros (Cox and Oakes (1984)) e de fácil interpretação. Além disso, de acordo com Nardi and Schemper (2003), uma boa discriminação entre modelos paramétricos exige que a censura não exceda 40-50 por cento dos casos. Os nossos dados são de dimensão moderada e cerca de 85 por cento dos casos correspondem a tempos de censura. No entanto, para efeitos de comparação, considerou-se a possibilidade de adaptar modelos paramétricos em trabalhos futuros, uma vez que podem ter um melhor desempenho do que o modelo Cox, se as condições exigidas forem cumpridas.

Como referido anteriormente, uma grande proporção de indivíduos não teve hospitalizações apesar do seu longo historial de doença, sugerindo que há indivíduos que nunca irão estar sujeitos ao evento de interesse, mesmo que o tempo sob observação possa ser prolongado indefinidamente. Por conseguinte, considerou-se a possível existência de indivíduos ”curados para a hospitalização”. Para explorar esta hipótese, foram implementados modelos de cura considerando o mesmo conjunto de covariáveis, para efeitos de comparação (Cai et al. (2012)). Pressupõe-se que os indivíduos provêm de uma mistura de duas populações - aqueles que irão experimentar uma hospitalização e aqueles que não irão. Para além de produzir uma estimativa da probabilidade de ”cura”, espera-se obter uma estimativa mais realista da função de sobrevivência para a fração da população que irá sofrer o evento. Assim, para estudar as hospitalizações, foram considerados 93 pacientes onde 12 deles apresentaram a necessidade de hospitalização ao longo do tempo de seguimento da doença. Os restantes 81 tratam-se de doentes que desde que foram diagnosticados com colite ulcerosa, não apresentaram a necessidade de serem hospitalizados. Para o estudo dos modelos de Cox e para responder aos objetivos propostos, foram criados modelos que combinavam variáveis como a Extensão, a idade ao diagnóstico, o Nancy score, o Mayo score, a presença ou não de envolvimento para além da margem endoscópica e a severidade histológica. A idade foi a única variável que se apresenta significativamente associada com a necessidade de um indivíduo ser hospitalizado. À medida que um indivíduo avança na idade, o risco de ser hospitalizado diminui, considerando-se assim um fator protetor para o desfecho em questão. Como referido anteriormente, devido à elevada percentagem de censuras neste desfecho, e como estas tendem a acontecer num período já avançado da doença, colocou-se a hipótese de os indivíduos, independentemente do tempo que estariam em observação, não iriam necessitar de ser hospitalizados. Deste acontecimento, surgiu a hipótese de, na população em estudo, haver a mistura de dois tipos de população – indivíduos suscetíveis à hospitalização e indivíduos não suscetíveis. Neste tipo de modelos, há duas componentes que devem ser tomadas em atenção – uma componente que define a suscetibilidade dos indivíduos, e outra que define o tempo de sobrevivência dos indivíduos suscetíveis. Quanto à suscetibilidade, foi possível observar que a variável Nancy score é a única que influencia esta componente. Um nível mais elevado desta variável favorece a suscetibilidade de um indivíduo face a outro com um nível mais baixo. Quanto à componente de tempo até à hospitalização, nenhuma variável é significativa, portanto nenhuma variável influencia significativamente o tempo até à ocorrência de uma hospitalização de entre os indivíduos suscetíveis.

Estas metodologias, apesar de diferentes, têm uma possível comparação. A componente de falha tem em conta a sobrevivência dos indivíduos suscetíveis e estes resultados podem ser comparados com os do modelo de Cox, que estuda a sobrevivência dos indivíduos. O que os modelos de cura permitem é considerar “pesos” diferentes para os indivíduos, e assim proporcionar conclusões mais fidedignas acerca dos tempos de sobrevivência para os indivíduos que vão sofrer o evento. Futuramente, muito mais poderá ser feito. Este estudo foi, essencialmente, um estudo exploratório, pois a dimensão da amostra não é a ideal. No entanto, apesar dos resultados não apresentarem grande evidência estatística, foi possível aplicar metodologias que abordassem a elevada percentagem de censuras que este estudo apresenta, representando um desafio do ponto de vista académico.

Palavras-chave: Modelos de Cox, Modelos de Cura, Colite Ulcerosa

Abstract

Ulcerative colitis is an idiopathic chronic inflammatory bowel disease characterized by periods of remission and acute exacerbations, and unpredictable disease course. Despite the availability of new drugs, there are still a considerable proportion of patients that will be refractory to therapy, need colectomy or develop complications, highlighting the need for better predictive markers of disease course (Bryant et al. (2016)).

Considering the endoscopic extent of the disease at the time of diagnosis, patients were divided into three major subgroups: inflammation limited to the rectum (E1), left-sided colitis with inflammation distally to the splenic flexure (E2) and extensive colitis with involvement proximal to the splenic flexure (E3). The combination of endoscopy and histology seems to provide a better indication of disease activity than endoscopy alone, especially in endoscopically non-inflamed mucosa. No study has assessed the prognostic value of histological features at diagnosis, and therefore it remains unknown whether the pre-treatment microscopic findings can predict the course of the disease. This research topic was proposed by two gastroenterology doctors (Catarina Frias Gomes under the guidance of Dr. Joana Torres), who developed the clinical part of the project, as well as the study protocol. This dissertation then made it possible to answer the questions posed, developing the statistical component (under the guidance of professor Dr. Marília Antunes). Herein, we designed a multicenter retrospective study to evaluate the prognostic value of histologic features in treatment-naïve proctitis and left-sided ulcerative colitis patients. The goals of this study were to evaluate the prevalence of microscopic inflammation in the macroscopically uninflamed mucosa, as well as the predictive value of the histological features of disease (in the endoscopically inflamed and non-inflamed mucosa) at the time of diagnosis. In Portugal there has been a considerable consumption of medical resources associated to UC. Between 2010 and 2017, around 3400 people were hospitalised due to this disease, with an average hospital stay of 10 days with a minimum of 0 days and a maximum of 109 days, making the study of hospitalisations due to this disease a very relevant issue in order to have a perspective of its occurrence, as well as to help in the monitoring of patients and in clinical and hospital administration decision making. In this work, we analyse the time to the need of hospitalisation of 93 patients diagnosed with UC in stages E1 and E2.

Cox models were used since proportional hazards assumption was not rejected and Cox models are popular amongst medical researchers, especially concerning the parameter interpretation (Cox and Oakes (1984)). Moreover, according to (Nardi and Schemper (2003)), a good discrimination among parametric models require censoring not to exceed 40-50 per cent. Our data is of moderate size and about 85% of the cases correspond to censored times. Nevertheless, for comparison purposes, we consider fitting parametric models in future work since they can perform better than the Cox model if the required conditions are met.

As said above, a large proportion of individuals had no hospitalisations despite their long disease history, suggesting that there are individuals who will never experience the event of interest, even if the time under observation could be prolonged indefinitely. Therefore, we considered the possible existence of “healed to hospitalisation” individuals. To explore this hypothesis, cure models were implemented considering the same set of covariates for comparison purposes (Cai et al. (2012)). It is assumed that the individuals come from a mixture of two populations – those who will experience a hospitalisation and those who will not. Besides producing an estimate of the probability of “cure”, it is expected to obtain a more realistic estimate of the survival function for the fraction of the population who will suffer the event.

For the study of this outcome, all the 93 patients were considered, twelve of which had been hospitalised at some time and the remaining were considered censored. Variables age, extent of the disease to diagnosis, endoscopic severity to diagnosis, histological severity and involvement beyond the endoscopic margin were included as covariates. In the Cox model, only age was significant, with older individuals being expected to have later hospitalisations. After the study of the Cox models, and given the high percentage of censorships, the hypothesis that individuals who had not suffered the event until then would probably not suffer it even if the study was prolonged was raised. Thus, with interest in studying whether the censored individuals might be non susceptible to hospitalisations or whether they would suffer the event later, the idea of studying the cure models emerged.

One of the main interests was to calculate what percentage of the population was non-susceptible, as well as to verify which variables could influence this choice. In the component defining the occurrence of cure, the only relevant variable was Nancy score, with higher level of Nancy score favouring susceptibility. None of the variables showed relevance to describe the failure time.

Much can be improved as future work if one of the major limitations of this study, the sample size, is overcome. Due to the fact that the sample is small and the events of interest are infrequent, the results obtained throughout the study were not statistically significant and therefore could not be used to draw conclusions for clinical purposes nor to advise medical teams. However, since this is an exploratory study and as it is the first study in this area, the conclusions that are drawn (for example, which variables are good predictors for the occurrence of outcomes) can be used as a good indicator and starting point for future research.

Keywords: Cox Models, Cure Models, Ulcerative Colitis

Acknowledgments

I dedicate this work to my parents, Helena Almeida and Nelson Almeida, for the unconditional support that they provided me not only in my academic career, but in all the moments of my life. All the words of strength and courage keep me committed to my work. I also want to thank Ana Almeida, my sister and my role model. Her teaching and dedication inspire me daily and make me want to be better. Without them, none of this was possible.

I thank António Pereira and Afonso Pereira for their ability to make me smile in the most painful moments.

To Rúben Lavrador, I thank his patience, the ability to say the right words at the right time to teach me to trust myself more.

I thank my supervisor, Professor Marília Antunes, for all the knowledge she gave me, her patience and all the dedication she gave me during this time.

I thank Dr. Joana Torres and Dr. Catarina Frias Gomes for their support and teaching. I take this opportunity to thank Hospital Beatriz Ângelo and the entire research team that contributed to the project, and thank them for the opportunity to use the medical database.

To Professor Lisete Sousa, Professor Tiago Marques, Professor Eduardo Severino, Ivo Sousa Ferreira, I want to thank you for all the time and help you have dedicated and contributed to the development of this project.

I would like to take this opportunity to thank all the professors of Departamento de Estatística e Investigação Operacional, of the Faculdade de Ciências da Universidade de Lisboa, and to Centro de Estatística e Aplicações de Universidade de Lisboa for helping me in my academic degree.

To my friends, who in one way or another contributed to my growth, thank you very much.

In particular to Catarina Queiroz, I thank you for all your support and friendship throughout my academic career. Six years have passed and we can say that we have learned a lot from each other. I hope it stays that way and that we can always share our achievements.

To Raquel Correia and Margarida Ribeiro I want to thank their friendship, understanding and patience.

Table of Contents

List of Tables

Chapter 1

Introduction

1.1 Motivation

"*Ulcerative colitis (UC) is an idiopathic, chronic inflammatory disorder of the colonic mucosa, which starts in the rectum and generally extends proximally in a continuous manner through part of, or the entire colon*" (Ordas et al. (2012)).

The cause of ulcerative colitis is unknown. Studies indicate that the inflammation in inflammatory bowel disease (IBD) involves a complex interaction of factors: the genes the person has inherited, the immune system, and something in the environment (Ordas et al. (2012)). Foreign substances (antigens) in the environment may be the direct cause of the inflammation, or they may stimulate the body's defenses to produce an inflammation that continues without control.

There are several symptoms, among which: abdominal pain, bloody diarrhea with mucus, fatigue or tiredness, weight loss, loss of appetite, anemia, elevated temperature, dehydration and feeling the urge to empty the bowels constantly (Basson (2015)). The seriousness of ulcerative colitis symptoms varies among affected people, over time and depends on which part of the bowel is affected.

There is no treatment hat can cure the disease in question, however, the UC can be controlled. Treatment will involve the use of different types of medications like: 5-aminosalicylic, steroids and biologics and also, in some cases, surgery. Long-standing ulcerative colitis can progress to colon cancer (Basson (2015)).

There are some risk factors associated with ulcerative colitis for example: age - more commonly affects people from the age of 15 years old to those in their 30s, ethnicity - caucasian people have a higher risk of developing the disease, especially those of Ashkenazi Jewish descent (from Eastern Europe and Russia), genetics - people with a close relative with ulcerative colitis have a higher risk of developing the disease and the use of Isotretinoin (Accutane) - this medication is sometimes used to treat severe acne.

"*North America and northern Europe have the highest incidence and prevalence rates of ulcerative colitis, with incidence varying from nine to 20 cases per 100000 person-years, and prevalence rates from 156 to 291 cases per 100000 people. Rates are lowest in the southern hemisphere and eastern countries (appendix pp 1–2). Incidence has increased in countries that have adopted an industrialised lifestyle, which suggests that environmental factors might be crucial in the triggering of disease onset*" (Ordas et al. (2012)).

In Loftus (2004), the prevalence ranges from 37 to 246 cases per 100,000 persons for ulcerative colitis. In the portuguese population, in 2007, the estimated prevalence of UC was 71 per 100,000 (Rassen et al. (2011)).

Inflammatory bowel disease happens to about 16 new patients per 100,000 inhabitants per year, in

Europe. The growth of this number has occurred in the most industrialized countries due to bad eating habits, environmental factors and more sedentary lifestyles.

1.2 Objective

It is quite frequent to find histological inflammation above the endoscopic margin, but the prognostic value of this finding remains unknown, despite strong evidence suggesting that during follow-up histology is an important prognostic factor. For this, and to understand the prognostic value, if any, of the presence of histological inflammation above the endoscopic margin at the time of diagnosis in medium and long-term outcomes, it was gathered the endoscopic and pathology records at diagnosis both in the involved and in the uninvolved area. Biopsies beyond the endoscopic margin from the index colonoscopy were retrieved, and reviewed by an experienced gastrointestinal pathologist who were blinded to the clinical information.

If this hypothesis is correct, the microscopic features at diagnosis could supplement the endoscopic classification of disease extent. Moreover, histology could also help clinicians to determinate the best therapeutic regime for each patient, so that they can prevent a serious evolution of the disease.

Bearing in mind the objective, and since there are available the time until certain events occur, survival analysis models will be used. Once that some of these times are very long, the use of cure models was an approach to consider.

1.3 Overview

The overview of this document is as follows.

Chapter 2 provides details about the variables, explanatory and response variables, under study in order to meet the objective of the study. It also introduces some inclusion and exclusion criterias study-related.

Chapter 3 explains the basic theory for model formulation, estimation and diagnostics for Cox regression and Cure Mixture models, respectively.

Cox models were used since proportional hazards assumption was not rejected and Cox models are popular amongst medical researchers, especially concerning the parameter interpretation. Moreover, a good discrimination among parametric models require censoring not to exceed 40-50 per cent. Our data is of moderate size and about 85% of the cases correspond to censored times. The Cure Mixture models were implemented since there are a large proportion of individuals that never experience the event of interest. Therefore, we considered the possible existence of “healed to hospitalisation” individuals.

Chapter 4 describes the application of the models introduced in previous chapter.

Finally, chapter 5 summarizes the main conclusions from this work.

Chapter 2

Data Description

For this multicenter retrospective study ninety-three patients from thirteen different centers were included. There are thirty seven patients from Hvidovre Hospital, Copenhagen, thirty one from Master Dei Hospital, Malta, thirteen from Hospital Beatriz Ângelo, Loures, three from Hospital Garcia da Orta, Almada, one from Centro Hospitalar Universitário do Algarve, two from Centro Hospitalar de Setúbal, three from Hospital Prof. Dr. Fernando Fonseca, Amadora, two from Centro Hospitalar de Vila Nova de Gaia / Espinho and one from Hospital Senhora da Oliveira, Guimarães.

The aim of the study was to assess the impact of microscopic inflammation at diagnosis, in ulcerative colitis patients. Firstly, it was evaluated the microscopic inflammation severity of inflamed mucosa.

The `Nancy score` is a validated score to assess histologic severity in ulcerative colitis and it ranges from 0 to 4 (less severe to most severe respectively), and the grades are described below. At diagnosis, patients have chronic and acute infiltrates. Thus the patients were categorized in 2 groups based on the severity of acute infiltrates. The presence of chronic and mild acute features (`Nancy score` < 3) defined mild microscopic inflammation, whereas moderate to severe acute inflammation and ulcers (`Nancy score` ≥ 3) defined severe microscopic inflammation. Even though the presence of severe microscopic inflammation during UC follow-up is a well know prognostic factor, the impact of inflammation at diagnosis has not been studied. Secondly, the prevalence and predictive value of microscopic inflammation above the endoscopic margin is unknown. Thus, it was also evaluated whether the presence of microscopic inflammation above the margin was a prognostic factor.

The `Involvement beyond endoscopic margin` is a variable that reflects the presence of inflammation beyond what was seen endoscopically or not.

Finally, it was created a third group of patients, defined by those who has severe inflammation in inflamed mucosa (`Nancy score` ≥ 3) and, at the same time, microscopic inflammation above the endoscopic margin. Theoretically, these patients have the most severe microscopic features that may impact in prognosis. In this group, it is suppose to evaluate the overall predictive value of histology.

Various information

Various information is collected for each patient, so it is possible to answer some questions to achieve the purpose of this study. There are available some demographic variables at diagnosis, as date of birth, sex, race, family history of IBD (Inflammatory Bowel Disease), etc. There are also clinical variables, measured at diagnosis, such as extent of the disease at diagnosis (E_1 and E_2), disease phenotype (age at diagnosis, symptoms at diagnosis, extra-intestinal manifestations, need for treatment at diagnosis), and variables, measured in follow-up, such as time to first relapse, interval between relapses, endoscopic disease activity (`Mayo score`), maintenance therapy, number and location of biopsies per patient, duration of follow-up.

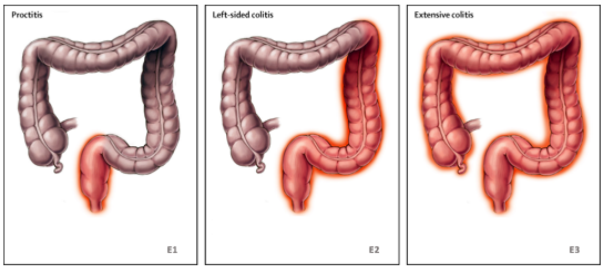


Figure 2.1: Extent of the disease.

The variables that will be used for this study will be described below.

`Extent` is the variable that measures the extent of the disease at the time of diagnosis. An E_1 patient is a patient who has limited inflammation to the rectum, an E_2 patient has more inflammation than the previous one, and can reach the left side of the colon and E_3 patient has inflammation in the entire colon. Figure 2.1 illustrates the three types of extension.

`Age` is the age of the patient at the time the disease was diagnosed.

`Mayo Score` assesses endoscopic inflammation severity in ulcerative colitis. This score ranges from 0 to 3, where 0 represents a normal mucosa appearance and corresponds to inactive disease and 3, the presence of ulcers and spontaneous bleeding. At diagnosis, every patient is categorized in one of these groups to assess endoscopic disease severity. Furthermore, the presence of deep ulcers and extensive ulcerations covering more than 10 % of the mucosal area are features of severe disease and are associated with worse prognosis.

The `Nancy score` and `Mayo score` were categorized in order to group individuals with higher and lower scores. The `Nancy score` was grouped as 0, 1 or 2 vs. 3 or 4, and the Mayo score as normal or mild disease vs. moderate or severe disease. The variable that combines the `Nancy score` and the `Involvement beyond endoscopic margin` will be categorized in two groups, for those who have histological severity (includes those who have `Nancy score` ≥ 3 and with `Involvement beyond endoscopic margin`) and those who don't (`Nancy score` < 3 with no `Involvement beyond endoscopic margin`, `Nancy score` < 3 with `Involvement beyond endoscopic margin` and `Nancy score` ≥ 3 and with no `Involvement beyond endoscopic margin`).

To evaluate the time until a particular event, times were created until each of the events considered, namely, use of steroids, hospitalisations, escalate therapy, acute severe colitis, colectomies and dysplasia.

2.1 Variables

The variables considered in this study are presented in Table 2.1, some of which were reorganized from the original database, others are variables that were created. The variables originally collected are shown in Attachment A.

Table 2.1: Description of each variable in database.

Variable	Description	Type	Categories
record_id	Identifies the patient	String	
dob	Date of birth	Date	
sex	Sex	Categorical Nominal	1- Male 2 - Female
diag_date	Date of diagnosis	Date	
diag_age	Age at diagnosis	Quantitive Continuous	
clinical_severity	Clinical severity at diagnosis	Categorical Nominal	1 - Mild UC 2 - Moderate UC 3 - Severe UC
extent	Extent at diagnosis	Categorical Nominal	1 - E1 Ulcerative proctitis 2 - E2 Left sided UC (distal UC)
severity	Endoscopic severity at diagnosis	Categorical Nominal	1 - Normal/ quiescent 2 - Mild disease 3 - Moderate disease 4 - Severe disease
ulceration	Presence of deep and extensive ulcerations	Categorical Nominal	1 - Yes 0 - No
followup_date	Date of the last follow-up	Date	
followup_days	Days of follow-up	Quantitive Continuous	
hospit_followup	Need for hospitalisation at diagnosis or during the follow-up	Categorical Nominal	1 - Yes 0 - No
hospit_number	Number of hospitalisations	Quantitive Discrete	
hospit_1	Date of 1st hospitalisation	Date	
hospit_2	Date of 2nd hospitalisation	Date	
hospit_3	Date of 3rd hospitalisation	Date	

Variable	Description	Type	Categories
hospit_4	Date of 4th hospitalisation	Date	
hospit_5	Date of 5th hospitalisation	Date	
hospit_6	Date of 6th hospitalisation	Date	
hospit_7	Date of 7th hospitalisation	Date	
hospit_8	Date of 8th hospitalisation	Date	
colectomy	Colectomy at diagnosis or during follow-up	Categorical Nominal	1 - Yes 0 - No
colectomy_date	Date of colectomy	Date	
asuc	Acute severe ulcerative colitis at diagnosis	Categorical Nominal	1 - Yes 0 - No
asuc_date	Date of acute severe ulcerative colitis at diagnosis	Date	
proximal_doc	Proximal extension of the disease	Categorical Nominal	1 - Yes 0 - No
proximal_doc_spec	Specify the proximal extension according to Montreal Classification	Categorical Nominal	1 - E1 to E2 2 - E1 to E3 3 - E2 to E3
steroids	Steroids used during follow-up	Categorical Nominal	1 - Yes 0 - No
steroids_number	Number of courses of steroids used during follow-up	Quantitive discrete	
steroids_date	Date of first flare requiring steroids	Date	
escalate_thr	Need to escalate therapy during follow-up	Categorical Nominal	1 - Yes 0 - No

Variable	Description	Type	Categories
escalate_thr_5asa	Increase dose of 5-ASA agents	Categorical Nominal	1 - Yes 0 - No
escalate_thr_5asa_date	Date of increase dose of 5-ASA	Date	
escalate_thr_azat	Azathioprine/ 6-MP	Categorical Nominal	1 - Yes 0 - No
escalate_thr_azat_date	Date of Azathioprine/ 6-MP	Date	
escalate_thr_cyclo	Cyclosporine	Categorical Nominal	1- Yes 0 - No
escalate_thr_cyclo_date	Date of cyclosporine	Date	
escalate_thr_metho	Methotrexate	Categorical Nominal	1 - Yes 0 - No
escalate_thr_metho_date	Date of Methotrexate	Date	
escalate_thr_bio	Biologics (anti-TNF or vedolizumab)	Categorical Nominal	1 - Yes 0 - No
escalate_thr_bio_date	Date of biologics	Date	
followup_dysplasia	Colonic mucosal dysplasia evidenced during the follow-up	Categorical Nominal	1 - Yes 0 - No
followup_dysplasia_spec	Specify type of dysplasia	Categorical Nominal	1 - Low grade dysplasia 2 - High grade dysplasia 3 - Colorectal cancer
followup_dysplasia_date	Date of dysplasia	Date	
Involvement	Involvement beyond endoscopic margin	Categorical Nominal	1 - Yes 0 - No

Variable	Description	Type	Categories
Nancy	Nancy score of the inflamed part	Categorical Ordinal	0 1 2 3 4
Mayo	Endoscopic severity at diagnosis	Categorical Ordinal	1 - Normal/quiescent 2 - Mild disease 3 - Moderate disease 4 - Severe disease
Histological severity	Combination of Nancy score and involvement beyond endoscopic margin	Categorical Nominal	1 - Nancy $\geq$ 3 with involvement 0 - Otherwise

2.2 Inclusion and Exclusion Criteria

In this study will be included all patients from different medical centers from 2006 to 2016, that regard the following criterias: patients diagnosed with ulcerative colitis, more specifically with proctitis or left sided colitis; people from different age groups (pediatric and adults); the patients must have at least two different colonic segments available for review and all of their information (endoscopic and histological) must be available in their clinical process. Proctitis, E_1 is defined as macroscopic inflammation not extending beyond the recto-sigmoid junction (15 cm from anus) and E_2 is defined as macroscopic inflammation not extending beyond the splenic flexure.

All patients that were diagnosed with pancolitis; those that don't have a detailed clinical information during follow-up; and those whose biopsies were made after starting therapy were excluded.

Chapter 3

Survival Analysis Models

In survival analysis, interest centres on a group or groups of individuals for each of whom (or which) there is defined a point event, often called failure, occurring after a length of time called the failure time. Failure can occur at most once on any individual (Cox and Oakes (1984)).

In this context the main interest is the study of survival times and of the factors that influence them. Types of studies with survival outcomes include clinical trials, prospective and retrospective observational studies, and animal experiments (Moore (2016)).

Survival times may arise from various contexts, such as biomedical sciences, economy, physics, engineering, psychology, demography (Rocha and Papoila (2009)). Examples of survival times, in the health context, include time from birth until death, time from entry into a clinical trial until death or disease progression, or time from birth to development of breast cancer (that is, age of onset). Survival studies can involve estimation of the survival distribution, comparisons of the survival distributions of various treatments or interventions, or elucidation of the factors that influence survival times (Moore (2016)).

3.1 Basic Concepts

- **Life time**

  Time elapsed from the initial well-defined instant to the occurrence of an event of interest.

- **Survival data**

  Data that represent the life time or survival time of individuals belonging to a given population.

- **Censored data**

  In essence, censoring occurs when we have some information about individual survival time, but we don't know the survival time exactly. There are generally three reasons why censoring may occur:

  1. a person does not experience the event before the study ends;
  2. a person is lost to follow-up during the study period;
  3. a person withdraws from the study because of death (if death is not the event of interest) or some other reason (e.g., adverse drug reaction or other competing risk).

(Kleinbaum and Klein (2012))

There are three types of censored data: right, left and interval. In this study, the data are right censored. Right censored data is identified when a lifetime exceeds a certain value, since the observation of the individual ends before the occurrence of the event of interest.

- **Truncation**

  There is truncation when only those individuals to whom a certain event occurs are studied, due to a certain selection process associated to the study. The remaining individuals who do not present a certain condition are hidden from the study.

  The most frequent type of truncation is left truncation, that arises when individuals come under observation only some known time after the natural time origin of the phenomenon under study (Cox and Oakes (1984)).

  In our study, truncation does not occur.

Survival and Hazard Functions

The actual survival time of an individual, t, can be regarded as the value of a variable T (Collett (2003)). This random variable T, must take non-negative values and must be continuos. The variable represents the life time of an individual belonging to a homogeneous population. Therefore, individuals do not differ from each other in relation to factors that may influence their survival (Rocha and Papoila (2009)).

The different values that T, the random variable associated with the survival time, can take have a probability distribution with underlying probability density function $f(t)$. The distribution function is then given by

$$F(t) = P(T < t) = \int_0^t f(u)du \tag{3.1}$$

and represents the probability that the survival time is less than some value t.

Survival Function

The survival function at time t, $S(t)$, is defined as the probability of the individual survives beyond the time t

$$S(t) = 1 - F(t) = P(T \geq t), \quad t \geq 0. \tag{3.2}$$

It is a non-increasing monotonous function such that

1. $S(0) = 1$;
2. $S(+\infty) = \lim_{t \to +\infty} S(t) = 0$

The probability density function is given by $f(t) = -S'(t)$.

Hazard Function

The hazard function is widely used to express the risk or hazard of death at some time t, conditional on he or she having survived to that time (Collett (2003)). The hazard may be regarded as the changing rate of the conditional probability of dying at time t given the survival time is no less than t. Because of this interpretation, $h(t)$ sometimes is also called instantaneous failure rate (Li and Ma (2013)).

The hazard function is defined as

$$h(t) = \lim_{\delta t \to 0^+} \frac{P(t \leq T < t + \delta t \mid T \geq t)}{\delta t}. \tag{3.3}$$

From (3.3), $h(t)\delta t$ is the approximate probability that an individual dies in the interval $(t, t + \delta t)$, conditional on that person having survived to time t (Collett (2003)).

The hazard function has the following properties

1. $h(t) \geq 0$;
2. $\int_0^{+\infty} h(t)\,dt = +\infty$.

From the expression (3.3), it is possible to obtain some relationships between survival and hazard functions. As $f(t) = -S^{'}(t)$ we have

$$h(t) = \frac{f(t)}{S(t)} = -\frac{S^{'}(t)}{S(t)} = -\frac{d\log S(t)}{dt}. \tag{3.4}$$

The survival probability $S(t)$ can be expressed as the inverse function of Equation (3.4) and since $S(0) = 1$,

$$\begin{aligned} S(t) &= \exp\left(-\int_0^t h(u)\,du\right) \\ &= \exp(-H(t)), \quad t \geq 0 \end{aligned} \tag{3.5}$$

where $H(t)$ is the integration of all hazard rates from time 0 to t, defined as the continuous cumulative hazard function at time t.

Similarly, from expression (3.5), the cumulative hazard function $H(t)$ can be expressed in terms of $S(t)$, given

$$H(t) = \int_0^t h(u)\,du, \quad t \geq 0. \tag{3.6}$$

Function $H(t)$ is a non-negative and monotonous increasing function that will be

1. linear if $h(t)$ is constant
2. convex if $h(t)$ is increasing
3. concave if $h(t)$ is decreasing

(Rocha and Papoila (2009))

Furthermore, from Equations (3.4) and (3.5), the probability density function $f(t)$ can be written in terms of the hazard function:

$$f(t) = h(t)\exp\left(-\int_0^t h(u)du\right). \tag{3.7}$$

(Liu (2012))

The hazard function describes the evolution over time of the instantaneous death probability of an individual. It represents, therefore, an aspect of the distribution of the time of life that has immediate physical meaning.

In fact, the hazard function can be

- **Monotonous increasing:** individuals are observed in a period of their life during which gradual aging occurs, i.e. the proportion of individuals dying at a given time, from among survivors at that time, increases with time.
- **Monotonous decreasing:** the longer the individual survives, the lower the probability of death at the subsequent time
- **Constant:** univocally characterizes the life time as exponentially distributed
- **Bathtub-shaped:** occurs in populations in which individuals are followed from birth to actual death, whether populations of living beings or manufactured objects
- **Hump-shaped:** the risk of death is initially increasing and decreases after some time

Parametric Models

A number of parametric models for survival times have been proposed in the literature, in particular in the context of reliability studies of technical equipment (Aalen et al. (2008)). The most common distributions are presented below.

Exponencial distribution

This distribution is characterized by having a constant hazard function. Exponencial distribution depends only on one parameter $\lambda > 0$.

$$\begin{aligned} &h(t;\lambda) = \lambda \\ &S(t;\lambda) = \exp(-\lambda t), \quad t \geq 0. \end{aligned} \tag{3.8}$$

Weibull distribution

Weibull is the most commonly parametric model, used in survival analysis, with two parameters. A parameter of scale λ and a shape parameter γ, both strictly positive.

$$\begin{aligned} &S(t;\lambda,\gamma) = \exp(-\lambda t^{\gamma}) \\ &h(t;\lambda,\gamma) = \lambda\gamma t^{\gamma-1} \\ &f(t;\lambda,\gamma) = \lambda\gamma t^{\gamma-1}\exp(-\lambda t^{\gamma}), \quad t \geq 0. \end{aligned} \tag{3.9}$$

The hazard function can be

- Monotonous increasing if $\gamma > 1$
- Monotonous decreasing if $0 < \gamma < 1$
- Constant if $\gamma = 1$ (exponencial distribution)

Gamma distribution

The gamma distribution depends on the scale parameter $\lambda > 0$ and the shape parameter $\alpha > 0$.

The density and hazard functions are

$$f(t;\lambda,\alpha) = \frac{\lambda^{\alpha}}{\Gamma(\alpha)} t^{\alpha-1}\exp(-\lambda t) \quad \text{and}$$
$$h(t;\lambda,\alpha) = \frac{\lambda^{\alpha}}{\Gamma(\alpha,\lambda t)} t^{\alpha-1}\exp(-\lambda t) \tag{3.10}$$

respectively, defined for $t > 0$, where $\Gamma(\alpha,x) = \int_x^{+\infty} u^{\alpha-1}e^{-u}\,du$ is the (upper) incomplete gamma function (Aalen et al. (2008)).

The hazard function can be

- monotonous increasing if $\alpha > 1$, with $h(0) = 0$ and $\lim_{t\to+\infty} h(t) = \lambda$
- monotonous decreasing if $0 < \alpha < 1$, with $\lim_{t\to 0^+} h(t) = +\infty$ and $\lim_{t\to+\infty} h(t) = \lambda$
- constant if $\alpha = 1$ (exponencial distribution)

Gompertz distribution

For $t \geq 0$ and $\theta > 0$, the hazard and survival functions are given by

$$h(t;\theta,\alpha) = \theta\exp(\alpha t)$$
$$S(t;\theta,\alpha) = \exp\left\{\frac{\theta}{\alpha}(1-\exp(\alpha t))\right\}. \tag{3.11}$$

The hazard function is monotonous increasing when $\alpha > 0$; for $\alpha < 0$, the hazard function is monotonous decreasing and the distribution function is improper, since $F(+\infty) < 1$ which is equivalent to $S(+\infty) > 0$.

Nonparametric Estimation

The survival function $S(t) = P(T \geq t)$, measures the probability of observing a survival time longer than a fixed value t, as said before. In many medical studies, it is not appropriate to assume any parametric form such as normal for the distribution of T. Therefore, the estimation of $S(t)$ cannot be simplified to an estimation problem with a finite number of unknown parameters. A nonparametric or distribution-free method has to be used to determine $S(t)$ for all possible t (Li and Ma (2013)).

Estimation of Survival Distribution - Kaplan-Meier Estimator

The well-known Kaplan-Meier estimator will be presented for the survival function, which is undoubtedly the most used in clinical studies and has been gaining more space in reliability studies (Colosimo and Giolo (2006)).

Let

$t_{(1)},\ldots,t_{(r)}$ be the different death instants in a sample of size n $(r \leq n)$.

d_i the number of deaths ocorred in $t_{(i)}$, $i = 1, ..., r$
and n_i the number of subjects in risk, in the study immediately before $t_{(i)}$, $i = 1, ..., r$.
The empirical survival function for censored data is

$$\hat{S}(t) = \prod_{i:t_{(i)} \leq t} \left(1 - \frac{d_i}{n_i}\right). \tag{3.12}$$

$\hat{S}(t) = 1$ for $0 \leq t < t_{(1)}$.
The variance of the Kaplan–Meier estimator is estimated by

$$\widehat{Var}[\hat{S}(t)] = [\hat{S}(t)]^2 \sum_{i:t_{(i)} \leq t} \frac{d_i}{n_i(n_i - d_i)}. \tag{3.13}$$

$\hat{S}(t)$ can be calculated recursively:

$$\begin{aligned} \hat{S}(t_{(1)}) &= 1 - \frac{d_1}{n_1} \\ \hat{S}(t_{(i)}) &= \hat{S}(t_{(i-1)}) \left(1 - \frac{d_i}{n_i}\right), \quad i = 1, ..., r. \end{aligned} \tag{3.14}$$

$\hat{S}(t)$ is a step function, with jumps at the observed death instants.
The Kaplan-Meier estimator has some properties such as:

- $\hat{S}(t) = 0$ for $t \geq t_{(r)}$, if $t_{(r)}$ is the largest registered observation, i.e., if the largest observation is not censored
- if the largest registered observation t^* is censored, the $\hat{S}(t)$ never reaches the value 0, and it is considered that the estimative is defined only in that instant, $\hat{S}(t) = \hat{S}(t_{(r)})$ for $t_{(r)} \leq t \leq t^*$
- $\hat{S}(t)$ is a consistent estimator of $S(t)$ and, under some regularity conditions, it can be considered the non parametric maximum likelihood estimator of $S(t)$
- the Kaplan-Meier estimator is self-consistent.

(Rocha and Papoila (2009))

3.2 Cox Regression

Let T be a continuous random variable that represents the lifetime.
For this model it is assumed that the hazard rate of an individual with covariates $z_1, ..., z_p$ takes the form

$$\begin{aligned} h(t; \mathbf{z}) &= h_0(t) \exp(\beta' \mathbf{z}) \\ &= h_0(t) \exp(\beta_1 z_1 + ... + \beta_p z_p). \end{aligned} \tag{3.15}$$

$\beta' = \beta_1, ..., \beta_p$ are the unknown regression coefficients that represents the effect of the covariates in the survival and $h_0(t)$ is a non-negative arbitrary function. The underlying hazard function $h_0(t)$ that represents the hazard function for an individual with vector $z = 0$.

This model effectively partitions the contribution of time and covariates to the hazard function into two multiplicative components:

The function $h_0(t)$, commonly referred to as the baseline function, specifies how the hazard function changes as a function of time when there is no covariate effect. The exponential term $\exp(Z'\beta)$ involves a simple linear combination of the covariates through coefficients β and does not depend on time. The coefficients β characterize the relative effects of different covariates on the function.

Interpretation of the coefficients

Consider two individuals with vectors of covariates z_1 and z_2, that differ only on the covariate z_j. So, the hazard function is given by

$$\begin{aligned}\frac{h(t;\mathbf{z}_1)}{h(t;\mathbf{z}_2)} &= \frac{h_0(t)\exp(\beta_1 z_{11} + \ldots + \beta_j z_{1j} + \ldots + \beta_p z_{1p})}{h_0(t)\exp(\beta_1 z_{21} + \ldots + \beta_j z_{2j} + \ldots + \beta_p z_{2p})} \\ &= \exp(\beta_j(z_{1j} - z_{2j})).\end{aligned} \tag{3.16}$$

Therefore, $\exp(\beta_j)$ represents the relative risk of occurrence of an event for two individuals that differ in a unit of the values of the covariable z_j, with the other values of covariates equal.

Fitting model (3.15) involves the complete determination of the unknown function $h_0(t)$ and the unknown parameters β. In practice knowing the parameter β can lead to a clear answer on the relationship between Z and survival outcome. The functional h_0 is thus treated as a nuisance and is only estimated if we want to draw the survival curve. To this end, we may employ a maximum likelihood estimation (MLE) approach to estimate β. Cox (1975) has shown that the estimation of β can be based on the following so-called partial likelihood function (Rocha and Papoila (2009)):

$$L(\beta) = \prod_{i=1}^{r} \frac{\exp(\beta'\mathbf{z}_{(i)})}{\sum_{l \in R_i} \exp(\beta'\mathbf{z}_{(l)})} \tag{3.17}$$

where $\mathbf{z}_i$ is the observed vector of covariates for the individual who dies at the i^{th} ordered death time t_i, $R_i = R(t_{(i)}) = \left\{j : t_j \geq t_{(i)}\right\}$ is the index set for all individuals who are alive and uncensored at a time prior to t_i, and the product is over all m distinct event times (Li and Ma (2013)).

This likelihood function does not depend on $h_0(t)$ and allows to do the inference about the parameter vector β. Note that this function is not a likelihood in the usual sense. Cox (1975) argued that it can be interpreted as a partial likelihood (intended to allow inference to be made in the presence of disturbing parameters, where $h_0(t)$ is a disturbing function) (Rocha and Papoila (2009)).

Under very general conditions, the partial maximum likelihood estimator of β is consistent, asymptotically normal with mean value β and covariance matrix $I(\beta)^{-1}$, where

$$I_{jk}(\beta) = -E\left(\frac{\partial^2 \log L}{\partial \beta_j \partial \beta_k}\right). \tag{3.18}$$

Covariate Selection

Sometimes it is necessary for investigators to agree on a proper model form. It is usual to identify which covariates have a significant influence on the survival of individuals, among all those that were considered.

The coefficient β_j represents the effect of the covariate z_j on an individual's survival. To assess whether there is evidence that this covariate significantly influences lifetime, we can test

$$H_0 : \beta_j = 0 \quad vs. \quad H_1 : \beta_j \neq 0.$$

The test statistic is of the form

$$Z = \frac{\hat{\beta}_j}{s.e.(\hat{\beta}_j)} \sim N(0,1) \tag{3.19}$$

where "s.e." stands for "standard error" (Moore (2016)).

Equivalently we have the following test statistic,

$$Q = \frac{\hat{\beta}_j^{\,2}}{\widehat{Var}(\hat{\beta}_j)} \sim \chi^2_1. \tag{3.20}$$

Note that the hypothesis that the covariate z_j has no significant influence on survival in the presence of the remaining covariates is being tested.

In general, the estimates $\hat{\beta}_j$ are correlated, which makes it difficult to interpret test results on the covariate coefficients included in a model. It is therefore preferable to resort to methods that allow comparison of alternative models.

More generally, the importance of one or more parameters may be tested using the likelihood ratio test, based on the deviance statistic.

Consider a model M_1 given by $h_0(t)\exp(\beta_1 z_1 + ... + \beta_p z_p)$ and a second model M_2 $h_0(t)\exp(\beta_1 z_1 + ... + \beta_p z_p + \beta_{p+1} z_{p+1} + ... + \beta_{p+q} z_{p+q})$, where the q terms are added in M_2.

If the terms added to the model are useful to explain the response variable, they will be maintained. If not, they can be omitted and model 1 is the most suitable of the two.

The deviance statistic compares the likelihood functions of the adjusted model M_1 to the model M_2 and is calculated as follows:

$$H_0 : \beta_{p+1} = ... = \beta_{p+q} \quad vs. \quad H_1 : \exists i : \beta_i \neq 0, i = p+1, ..., p+q.$$

Under the null hypothesis,

$$-2\ln(\hat{L}_1/\hat{L}_2) \sim \chi^2_q \tag{3.21}$$

where $\hat{L}$ is the maximized likelihood function for each model. Under the null hypothesis that the extra q parameters in model 2 are all equal to zero, the test statistic has an asymptotic chi-square distribution with q degrees of freedom.

Testing the hypothesis of proportional hazards

In the Cox model the following

$$\log[-\log S(t;\mathbf{z})] = \beta'\mathbf{z} + \log[-\log S_0(t)]. \quad (3.22)$$

The validity of the proportional hazards hypothesis is often explored for each covariate separately. For a covariate with m categories, consider the m groups made up of individuals with equal values of this covariate.

A graphical evaluation of this hypothesis consists in obtaining the Kaplan-Meier estimate of survival function in each of these groups of individuals and to represent $\log[-\log \hat{S}_i(t)]$ vs. t on the same graph for $j = 1,...,m$.

If the plots are reasonably parallel, it seems plausible to admit the proportionality of the risk functions.

Let $\mathbf{z} = (z_1,...,z_p)'$ be the covariate vector for those it is suppose to study the hypothesis of proportional hazards.

Consider covariate z_j and let $\mathbf{z} = (z_j, \mathbf{z}^-)$ where $\mathbf{z}^- = (z_1,...,z_{j-1},z_{j+1},...,z_p)$. Considering that z_j is dichotomous, assuming the values 0 and 1, we will have two strata. The hazard function will be given by

$$h(t;\mathbf{z}) = \begin{cases} h_{01}(t)\exp(\beta'^{-}\mathbf{z}^-) & \text{if } z_j = 0 \\ h_{02}(t)\exp(\beta'^{-}\mathbf{z}^-) & \text{if } z_j = 1. \end{cases} \quad (3.23)$$

It is supposed that the variables included in z^- satisfy the hypothesis of proportional hazards and the estimate of β^- is obtained by maximizing partial likelihood over the entire sample.

The underlying survival functions are estimated separately in each stratum using the estimation $\hat{S}_0(t)$ or $\tilde{S}_0(t)$

$$\tilde{S}_0(t) = \prod_{i:t_{(i)} \le t} \left[\exp\left(-\frac{d_i}{\sum_{l \in R_i} \exp(\hat{\beta}\mathbf{z}_l)} \right) \right]. \quad (3.24)$$

A graphical representation of $\log[-\log \hat{S}_{01}(t)]$ and $\log[-\log \hat{S}_{02}(t)]$ versus t is made. If the plots are reasonably parallel, the covariate z_j can be included in Cox model.

If variable z_j has m categories, then it is necessary to estimate m survival functions after excluding the corresponding $m-1$ indicator variables.

In order to test the hypothesis of proportional hazards for the fixed covariate z_j, a time-dependent variable is proposed. This new variable has the following form $z_j(t) = z_j g(t)$ and is included in the model that has p covariates:

$$h(t;\mathbf{z}) = h_0(t)\exp(\beta_j z_j + \gamma z_j(t) + \beta^- \mathbf{z}^-) \quad (3.25)$$

where $\mathbf{z}^- = (z_1,...,z_{j-1},z_{j+1},...,z_p)$. When $\gamma \neq 0$ it means that there is a variation in time in the risk function ratio for two individuals with different values of z_j. The $g(t)$ function is usually used for simple functions such as $g(t) = t$ or $g(t) = \log t$.

In order to test $H_0 : \gamma = 0$ *vs.* $H_1 : \gamma \neq 0$ it is used the Wald test or the likelihood ratio test.

3.3 Residual Analysis

Several types of residuals have been defined and used for model checking in linear and generalized linear models. Many of these have been generalized to survival analysis. In addition, the fact that survival data evolves over time, and requires special assumptions such as proportional hazards, makes it necessary to develop additional diagnostic residual methods (Moore (2016)).

Cox-Snell Residuals

The Cox–Snell residual is originally developed to define residuals associated with different model functions, including the log-linear regression (Liu (2012)). These residuals are used to evaluate the global adjustment of the final model. Consider the distribution and survival functions as uniform (0,1). As $H(T) = -\log S(T)$, $H(T)$ has an unit exponencial distribution.

Since in the Cox model, we have

$$H(t;z) = \exp(\beta' z)H_0(t) \tag{3.26}$$

the residual for the i-th individual, $i = 1, \ldots, n$ is given by

$$r_i = \hat{H}(t_i;\mathbf{z}_i) = \exp(\hat{\beta}' z_i)\hat{H}_0(t_i). \tag{3.27}$$

If the model adjusted to the data is appropriate, then the estimated values $\hat{H}(t_i;z_i)$ will have equal properties to the actual values.

Schoenfeld residuals

Schoenfeld residuals, are defined specifically for the proportional hazard model, with computations performed within the context of the Cox model. Like the specification of the partial likelihood, the derivation of the Schoenfeld residuals does not depend on time but, rather, on the rank order of survival times (Liu (2012)).

For the i-th individual, the Schoenfeld residual correspondent to the z_j covariate, $j = 1, ..., p$ is given by

$$r_{ji} = \delta_i(z_{ji} - a_{ji}) \tag{3.28}$$

where

$$a_{ji} = \frac{\sum_{l \in R_i} z_{jl} \exp(\beta' \mathbf{z}_l)}{\sum_{l \in R_i} \exp(\beta' \mathbf{z}_l)}. \tag{3.29}$$

For an individual whose death is observed in t_i, the residual is the difference between the value of z_j and the weighted average of this variable for all at-risk individuals in t_i. The weight associated with an individual $l \in R_i$ is $\exp(\hat{\beta}' \mathbf{z}_l)$.

In case the lifetime is not observed, the residuals are null and indicated as missing values.

$$\frac{\partial \log L}{\partial \beta_j} = \sum_{i=1}^{n} \delta_i \left\{ z_{ji} - \frac{\sum_{l \in R_i} z_{jl} \exp(\beta' \mathbf{z}_l)}{\sum_{l \in R_i} \exp(\beta' \mathbf{z}_l)} \right\} \tag{3.30}$$

where L is the partial likelihood function. The i-th portion of this sum, calculated in $\hat{\beta}$, is then the Schoenfeld residuals corresponding to the covariate z_j for the i-th individual.

The estimates $\hat{\beta}_j$ are given by

$$\left.\frac{\partial \log L}{\partial \beta_j}\right|_{\hat{\beta}} = 0. \tag{3.31}$$

(Rocha and Papoila (2009))

Because the partial likelihood estimator of the coefficient β, is the solution to the equations obtained by setting (3.30) equal to zero, the sum of the Schoenfeld residuals is zero (Hosmer et al. (2008)). The r_{ij} converges to 0 in probability as n tends to a large number (Liu (2012)).

Martingale residuals

The martingale residuals are useful in determining the functional form that should be used for a given covariable, in order to explain its effect on survival, as well as in the detection of outliers.

Observations with unusually large martingale residuals are not well fitted by the model (Collett (2003)). The martingale residual associated with the i^{th} individual is given by:

$$\hat{M}_i = \delta_i - \exp(\hat{\beta}'\mathbf{z}_i)\hat{H}_0(t_i)). \tag{3.32}$$

This residuals are skewed and take values in $(-\infty, 1)$, where the residuals corresponding to censored observations are negative.

Deviance residuals

Another type of residuals used for the identification of outliers are the deviance residuals. These are a transformation of martingale residuals being more symmetrically distributed around zero:

$$D_i = sgn(\hat{M}_i)\sqrt{-2 \times (\hat{M}_i + \delta_i \ln(\delta_i - \hat{M}_i))} \tag{3.33}$$

where the sign function ensures that the deviance residual for the i-th individual takes the same sign as the correspondent martingale residual.

3.4 Cure Models

As it was said before, the survival analysis consists of analyzing data from an initial moment to another where an event of interest occurs. However, it is not always possible to observe the event of interest, in which case the so-called censored observations arise.

However, if most of the censored observations correspond to the largest observations, we may suspect that there are individuals who will never experience the event of interest, even if the time under observation could be prolonged indefinitely (Abreu (2005)).

The mixture cure model is a special type of survival models. It assumes that the studied population is a mixture of susceptible individuals who may experience the event of interest, and immune/cure/non-susceptible individuals who will never experience the event (Cai et al.

(2012)). For individuals who are considered immune, the corresponding survival function will always take the value one at any time (degenerate survival function).

In a population we may have a mixture of immune (or non-susceptible or healed) individuals and non-immune (or susceptible or sick) individuals; For the non-immune it will correspond to a non-degenerate survival function.

Sometimes, a threshold value is defined and individuals who have lifetime values above this value are considered cured. Other times, the occurrence of a certain event (e.g. hospital discharge) distinct from the event of interest (e.g. death from the disease) will enable the identification of the healed individuals.

Mixture Cure Models

Let T represent the lifetime of an individual in a given population in which immune and susceptible individuals are assumed to exist and Y a binary random variable, where $Y = 1$ if the individual is susceptible and $Y = 0$ if the individual is immune.

Let $S(t|Y = 1)$ represent the survival function of the susceptible individuals and p denote the proportion of immune individuals in the population. The expression for the mixture cure model, at the expense of the survival function, is as follows:

$$S(t) = p + (1 - p)S(t|Y = 1). \tag{3.34}$$

Given the equality $h(t) = -\frac{\partial \log S(t)}{\partial t}$, the cure model in terms of the hazard function is defined as follows:

$$h(t) = \frac{(1 - p) f(t|Y = 1)}{S(t)}. \tag{3.35}$$

In 3.35, $S(t)$ is an improper survival function because $S(\infty) = p$, while $S(t|Y = 1)$ is a proper function. On the other hand, $h(t)$ has a finite integral in $(0,\infty)$ therefore it is not a hazard function in the usual sense.

In this type of cure models it is interesting to estimate not only the parameters of the distribution function of susceptible individuals, but also the proportion of immune individuals.

Non Mixture Cure Models

Since in a cure model the population survival function is an improper function and given the relationship between the survival function and the corresponding cumulative hazard function, the latter will be a superiorly limited function.

Thus, $S(t)$ is the survival function of T and $H(t)$ is such that $S(t) = \exp[-H(t)]$. If $S(\infty) > 0$, that exists $\theta < \infty$, such that $H(\infty) = \theta$. One possible way of characterizing this property is to consider $H(t) = \theta F(t)$, where $F(t)$ designates the (proper) distribution function of a nonnegative random variable. Thus, the non-mixture cure model can be written in the form

$$S(t) = \exp[-\theta F(t)], \tag{3.36}$$

the corresponding risk function is given by

$$h(t) = \theta f(t), \tag{3.37}$$

where $f(t)$ is the density function corresponding to $F(t)$. Thus, $S(\infty) = e^{-\theta}$ corresponds to the cure probability.

In this context, there are two hypotheses for obtaining an improper survival function. Either the parameter space of some proper distribution is modified, as happens, for example, with the modified Gompertz distribution, or it is considered a frailty variable to characterize the susceptibility of each individual to the event of interest.

3.4.1 Chen's Distribution

The cure model is a mixture model, in which the distribution of the lifetime of susceptible individuals is a biparametric distribution proposed by Chen (2000). This distribution was developed in the usual context of Survival Analysis, that is, without admitting the possibility of existence of cured individuals in the population under study. One of the merits of this model comes from the fact that the hazard function of susceptible individuals has some flexibility, in the sense that, according to the value of one of its parameters, it can be either monotonically increasing or bathtub shaped. Another advantage lies in the ease of mathematical treatment, as we shall see below, in the context of parameter estimation (Abreu (2005)).

The distribution function proposed by Chen is

$$F(t) = 1 - \exp\left[\lambda\left(1 - \exp(t^{\alpha})\right)\right], \quad t > 0 \tag{3.38}$$

where λ and α are the scale and shape parameters respectively. Thus, the survival and hazard functions are given by

$$\begin{aligned} S(t) &= 1 - F(t) = \exp\left[\lambda\left(1 - \exp(t^{\alpha})\right)\right], \quad t > 0 \\ \text{and} \quad h^*(t) &= \lambda \alpha t^{\alpha-1} \exp(t^{\alpha}), \quad t > 0 \quad \text{, respectively.} \end{aligned} \tag{3.39}$$

The function $h^*(t)$ is increasing when $\alpha \geq 1$, and, for $\alpha < 1$, is decreasing for $t \in \left[0, \left(\frac{1}{\alpha} - 1\right)^{\frac{1}{\alpha}}\right]$ and increasing for $t \geq \left(\frac{1}{\alpha} - 1\right)^{\frac{1}{\alpha}}$.

3.4.2 Cure Model based on Chen's Distribution

Considering the Chen's distribution for the lifetime distribution of susceptible individuals, the cure model, defined at the expense of the survival function is given by

$$S(t) = p + (1-p)\exp\left[\lambda\left(1 - \exp(t^{\alpha})\right)\right], \quad t > 0. \tag{3.40}$$

The model written at the expense of the hazard function is given by

$$h(t) = \frac{(1-p)\lambda \alpha t^{\alpha-1} \exp(t^{\alpha}) \exp\left[\lambda\left(1 - \exp(t^{\alpha})\right)\right]}{p + (1-p)\exp\left[\lambda\left(1 - \exp(t^{\alpha})\right)\right]}. \tag{3.41}$$

Although α is a shape parameter, the population hazard function is completely different from the hazard function of the susceptible individuals: $h(t)$ is decreasing for $\alpha \leq 1$ and unimodal for $\alpha > 1$. When $\alpha > 1$, the mode value will be bigger, the lower the value p, where p represents the immune individuals. λ, will define the value of t from which the hazard function will stabilize at a value close to zero, the value of t from which individuals surviving until that time can be expected to be healed. The smaller the λ value, the higher the t value from which the hazard function is approximately zero.

3.4.3 Parameter estimation

In this section, the aim is to find the parameter estimation thought maximization of the likelihood function.

If non-informative censoring and right censoring are admited, the likelihood function for a sample of dimension n may be expressed by

$$L = \prod_{i=1}^{n} f(t_i)^{\delta_i} S(t_i)^{1-\delta_i} \tag{3.42}$$

where δ_i is an indicator variable, such that $\delta_i = 1$ corresponds to a real lifetime and $\delta_i = 0$ to a censored observation.

Considering a sample of size n and $t_1,...,t_n$ the observed lifetimes (completed and censored), with no loss of generality, suppose that the first m $(m < n)$ lifetimes are censored. Let $\delta_1,...,\delta_n$ be such that

$$\delta_i = \begin{cases} 0 & \text{if} \quad 1 \leq i \leq m \\ 1 & \text{if} \quad m+1 \leq i \leq n \end{cases}$$

and $y_1,...,y_n$ such that

$$y_i = \begin{cases} 0 & \text{if the individual is immune} \\ 1 & \text{if the individual is susceptible.} \end{cases}$$

Considering the pair (δ_i, y_i), the type of contribution to the likelihood of the correspondent observation varies, as shown in the Table 3.1.

Table 3.1: Summary of the possibilities for $(\delta_i$,$y_i)$.

Possible values for (δ_i, y_i)	Description of the situation	Contribution to the likelihood
(0,0)	Censored, immune	p
(0,1)	Censored, susceptible	$(1-p)S(t\|\text{Y=1})$
(1,0)	Observed, immune	Impossible
(1,1)	Observed, susceptible	$(1-p)f(t\|\text{Y=1})$

When we have censored observations, it is not possible, as a general rule, to unequivocally identify which are the immune individuals. The fact that we do not observe all $y_i's$ (those corresponding to the censored lifetimes), leads us to a situation of incomplete data. In this specific case, the method of choice is the EM algorithm, since it is an iterative method that allows obtaining the maximum likelihood estimates of the parameters in situations in which there are omitted observations.

In the case of cure mixture model, the total observed likelihood is

$$L_O = \prod_{i=1}^{n} [(1-p)f(t_i \mid Y=1)]^{\delta_i} [p + (1-p)S(t_i \mid Y=1)]^{1-\delta_i}. \tag{3.43}$$

If all y_i were observed (usually not those that correspond to censored observations), the complete likelihood can be written

$$L_C = \prod_{i=1}^{n} [[(1-p)f(t_i \mid Y=1)]^{y_i}]^{\delta_i} \left[p^{1-y_i} [(1-p)S(t_i \mid Y=1)]^{y_i}\right]^{1-\delta_i}. \tag{3.44}$$

As $f(t \mid Y=1) = h(t \mid Y=1)S(t \mid Y=1)$ and given the possible values for (δ_i, y_i), $(1-\delta_i)(1-y_i) = 1-y_i$, the likelihood given in 3.44 can be factorized into

$$L_C = \prod_{i=1}^{n} (1-p)^{y_i} p^{1-y_i} \prod_{i=1}^{n} h(t_i \mid Y=1)^{y_i \delta_i} S(t_i \mid Y=1)^{y_i}. \tag{3.45}$$

Step E of this algorithm consists of determining the expected value of the complete likelihood logarithm in relation to the distribution of the unobserved values of Y, conditional on the actual parameter values and the observed data O, where O = {observed y_i, $(t_i, y_i), i = 1, ..., n$}. However, as with respect to the censored observations the logarithm of L_C is linear in Y, to calculate the expected value of log L_C simply replace in complete likelihood, the unobserved values of Y with the expected values denoted by t_i. The log-likelihood is given by

$$\begin{aligned} L_C = &\sum_{i=1}^{n} y_i \delta_i \log h(t_i \mid Y=1) + \sum_{i=1}^{n} y_i \log S(t_i \mid Y=1) + \\ &+ \sum_{i=1}^{n} y_i \log(1-p) + \sum_{i=1}^{n} (1-y_i) \log p. \end{aligned} \tag{3.46}$$

For the censored observations,

$$\begin{aligned} \tau_i = E(Y_i \mid O) &= P(Y_i = 1 \mid T_i > t_i, \delta_i = 0) \\ &= \frac{P(T_i > t_i, \delta_i = 0 \mid Y_i = 1) P(Y_i = 1)}{P(T_i > t_i, \delta_i = 0)} \\ &= \frac{(1-p)S(t_i \mid Y_i = 1)}{S(t_i)} \end{aligned} \tag{3.47}$$

where, Y is well described by a Bernoulli

$$\begin{cases} 0, & P(immune) \\ 1, & P(susceptible). \end{cases}$$

Thus, in complete likelihood function, each y_i is replaced by w_i, where w_i is defined as

$$w_i = \begin{cases} 1 & \text{if } \delta_i = 1 \\ \tau_i & \text{if } \delta_i = 0. \end{cases}$$

In step E, each censored observation i is assigned to subpopulation $Y=1$ with probability τ_i and to subpopulation $Y=0$ with probability $1-\tau_i$.

After replacing y_i with τ_i in the complete likelihood, the following "expected" likelihood is obtained

$$L_E = \prod_{i=1}^{n} q^{w_i}(1-q)^{1-w_i} \prod_{i=1}^{n} h(t_i \mid Y=1)^{w_i\delta_i} S(t_i \mid Y=1)^{w_i} = L_{E_1} L_{E_2} \qquad (3.48)$$

where $q = 1-p$.

Case of Chen's distribution

According to the definition of the expected values of the unobserved Y values, if the susceptible individuals have a Chen distribution,

$$\tau_i = \frac{q \times \exp\left[\lambda\left(1-\exp(t_i^{\alpha})\right)\right]}{1-q+q\times\exp\left[\lambda(1-\exp(t_i^{\alpha}))\right]}. \qquad (3.49)$$

Thus, since $w_i = 1$ if $\delta_i = 1$, and $w_i = \tau_i$ if $\delta_i = 0$, the said "expected" likelihood is

$$L_E = \prod_{i=1}^{n} q^{w_i}(1-q)^{1-w_i} \prod_{i=1}^{n} \left[\lambda \alpha t_i^{\alpha-1} \exp\left(t_i^{\alpha}\right)\right]^{w_i\delta_i} \left[\exp\left[\lambda\left(1-\exp\left(t_i^{\alpha}\right)\right)\right]\right]^{w_i} = L_{E_1} L_{E_2} \qquad (3.50)$$

and the logarithms of L_{E_1} and L_{E_2} are, respectively,

$$\begin{aligned} \log L_{E_1} &= (n-m)\log q + m\log(1-q) + (\log q - \log(1-q)) \sum_{i=1}^{m} \tau_i \\ \log L_{E_2} &= \lambda \sum_{i=1}^{m} \tau_i \left[1-\exp(t_i^{\alpha})\right] + (n-m)(\log\lambda + \log\alpha) + \\ &\quad + (\alpha-1) \sum_{i=m+1}^{n} \log t_i + \sum_{i=m+1}^{n} t_i^{\alpha} + \lambda \sum_{i=m+1}^{n} \left[1-\exp(t_i^{\alpha})\right] \end{aligned} \qquad (3.51)$$

bearing in mind that m censored observations are indexed from 1 to m and $n-m$ observed lifetimes are indexed from $m+1$ to n.

After deriving $\log L_{E_2}$, it is possible to get an explicit expression only for the estimator of λ

$$\hat{\lambda} = \frac{n-m}{\sum_{i=1}^{m} \tau_i \left[\exp\left(t_i^{\alpha}\right)-1\right] + \sum_{i=m+1}^{n} \left[\exp\left(t_i^{\alpha}\right)-1\right]}. \qquad (3.52)$$

The estimate of α is obtained using the Newton-Raphson method.

To estimate the parameters, that is, to use the EM algorithm, it is necessary to replace τ_i with the respective value, from which we obtain

$$\hat{q} = \frac{1}{n}\left[(n-m) + q \times \sum_{i=1}^{m} \frac{\exp\left[\lambda\left(1-\exp\left(t_i^{\alpha}\right)\right)\right]}{1-q+q\exp\left[\lambda\left(1-\exp\left(t_i^{\alpha}\right)\right)\right]}\right], \qquad (3.53)$$

$$\hat{\lambda} = \frac{n-m}{q \times \sum_{i=1}^{m} \frac{\exp\left[\lambda\left(1-\exp\left(t_i^{\alpha}\right)\right)\right]}{1-q+q\times\exp\left[\lambda\left(1-\exp\left(t_i^{\alpha}\right)\right)\right]} \left[\exp\left(t_i^{\alpha}\right)-1\right] + \sum_{i=m+1}^{n}\left[\exp\left(t_i^{\alpha}\right)-1\right]}. \tag{3.54}$$

In its current form, it not take into account the heterogeneity of the population under study. Information of the patients can be incorporated into the model in the form of independent variable such as in Cai et al. (2012). The notation addopted up to this point, is the same of Abreu (2005), is slightly different from the one adopted in Cai et al. (2012).

From now on, Cai et al. will be followed closely since it was used by the authors to build the package used in section 4, and thus, the analogy will be made.

Where Abreu (2005) has p as the proportion of immune individuals in the population, Cai et al. (2012) has this proportion as $1-\pi(\mathbf{z})$. This last one is the probability of a patient being cured depending on $\mathbf{z}$, and $S(t \mid \mathbf{x})$ the survival probability of uncured patients depending on $\mathbf{x}$, where $\mathbf{x}$ and $\mathbf{z}$ are observed values of two covariate vectores that may affect the survival function.

Thus, the mixture cure model, according to Cai et al. (2012), can be express as,

$$S(t \mid \mathbf{x},\mathbf{z}) = \pi(\mathbf{z})S(t \mid \mathbf{x}) + 1 - \pi(\mathbf{z}) \tag{3.55}$$

where $\pi(\mathbf{z})$ is referred to as ”incidence” and $S(t \mid \mathbf{x})$ is referred to as ”latency”. Thus, everything follows.

An advantage of the mixture cure models is that the proportion of cured subjects and the survival distribution of uncured subjects are modeled separately and the interpretation of effects of $\mathbf{x}$ and $\mathbf{z}$ is straightforward.

Usually, a logit link function

$$\pi(\mathbf{z}) = \frac{\exp(\mathbf{bz})}{1+\exp(\mathbf{bz})} \tag{3.56}$$

where $\mathbf{b}$ is a vector of unknown parameters, is used to model the effects of $\mathbf{z}$.

Chapter 4

Data Analysis

Ninety-three patients from thirteen different centers were included in this multicenter retrospective study.

4.1 Exploratory Analysis

In this section, the exploratory analysis of the study population is described in Table 4.1.

The continuous variables in this study are all time related, presented in years. For these variables, the mean, the minimum and the maximum, are presented. Regarding the nominal variables, the absolute and relative frequencies are presented.

Table 4.1: Main characteristics of study sample participants.

Baseline characteristics		
Sex (%)		
Male	51	(54.8)
Female	42	(45.2)
Age	44.03	(17-88)
Follow-up	4.6	(0.2-27.3)
Extent (%)		
E_1	61	(65.6)
E_2	32	(34.4)
Involvement beyond margin (%)		
Yes	66	(71.0)
No	27	(29.0)
Nancy score inflamed (%)		
0	2	(2.2)
1	6	(6.5)
2	13	(14.0)
3	55	(59.1)
4	17	(18.3)
Nancy score inflamed part (categorized) (%) *		
<3	21	(22.6)
≥ 3	72	(77.4)

Baseline characteristics		
Nancy score non-inflamed (%)		
0	52	(55.9)
1	3	(3.2)
2	10	(10.8)
3	1	(1.1)
NA	27	(29.0)
Mayo Score (%)		
Normal/quiescent	2	(2.2)
Mild disease	33	(35.9)
Moderate disease	44	(47.8)
Severe disease	14	(15.2)
Mayo Score (categorized) (%) *		
Low	35	(37.6)
High	58	(62.4)
Histological severity (%) *		
Yes	50	(53.8)
No	43	(46.2)

* The need to create this variable arises in chapter 4.2.

About 71% of the patients had presence of histological inflammation beyond the endoscopic margin, i.e, were misdiagnosed. The other 29% had no involvement beyond the endoscopic margin. This value corresponds to the percentage of NA (not available) present in the `Nancy score` of the non-inflamed part since, for this individuals, there is no inflamation beyond the endoscopic margin. Therefore, the score takes no value.

In order to assess the impact of the presence of histological inflammation and the histologic features of colonic mucosa beyond or proximally to the endoscopic margin on the medium and long-term outcomes, the time until the occurrence of each of the outcomes under analysis was calculated. The main features of these times are presented in Table 4.2. The time to the occurrence of the first event was also calculated. This variable corresponds to the time from the diagnosis to the occurrence of the first of a set of events. Thirty-nine patients were found to have experienced one event, and may or may not have had at least one event.

Table 4.2: Time until each event, in years.

Time to event	Minimum	Maximum	Mean
Steroids	0.04	12.18	2.17
Hospitalisation	0.02	3.31	1.34
Escalate therapy	0.08	5.96	2.05
Acute severe colitis	0.05	3.31	1.90
Colectomy	2.8	2.8	2.8
Dysplasia	0.12	6.27	3.19
First event	0.02	12.18	2.34

Since there is only one patient who had the need of colectomy, the minimum, maximum and mean time are the same. The event that corresponds to the smallest time to any event is an hospitalisation, whereas the largest time to any event corresponds to the need of steroids.

The Fisher exact test was used to evaluate the association between the occurrence of each of the six outcomes of interest and the variables `Involvement beyond endoscopic margin`, `Nancy score` and `Histological severity`. Table 4.3 presents the results of the Fisher exact test to the `Involvement beyond endoscopic margin`.

Table 4.3: Fisher exact test to evaluate the association of Involvement beyond the margin and each outcome.

	Involvement	No involvement	p-value
Steroids			
Yes	17	3	0.17
No	49	24	
Hospitalisation			
Yes	10	2	0.50
No	56	25	
Escalate Therapy			
Yes	11	1	0.17
No	55	26	
Acute severe colitis			
Yes	5	0	0.32
No	61	27	
Colectomy			
Yes	1	0	$\simeq 1$
No	65	27	
Dysplasia			
Yes	2	0	$\simeq 1$
No	64	27	

Although the p-values are none significant, it is possible to verify that, all the outcomes are more frequent when there is involvement beyond the endoscopic margin.

The same evaluation was made to the `Nancy score` in order to test if the `Nancy score` has some association in each of the outcomes.

Table 4.4: Fisher exact test to evaluate the association of Nancy score and each outcome.

	Nancy < 3	Nancy ≥ 3	p-value
Steroids			
Yes	3	17	0.55
No	18	55	
Hospitalisation			
Yes	1	11	0.29
No	20	61	
Escalate Therapy			
Yes	2	10	0.55

	Nancy < 3	Nancy ≥ 3	p-value
No	19	62	
Acute severe colitis			
Yes	0	5	0.58
No	21	67	
Colectomy			
Yes	0	1	$\simeq 1$
No	21	71	
Dysplasia			
Yes	1	1	0.40
No	20	71	

Again, all the p-values are none significant. It can be seen that most outcomes tend to occur more frequently when the score is ≥ 3, except in the case of dysplasia, where only two individuals present the outcome and are equally distributed among the scores.

Finally, it was performed some Fisher exact tests to evaluate the association of each outcome in the variable `Histological severity`. The results of this tests are presented in Table 4.5.

Table 4.5: Fisher exact test to evaluate the association of Histological severity and each outcome.

	Histological severity	No Histological severity	p-value
Steroids			
Yes	14	6	0.13
No	36	37	
Hospitalisation			
Yes	9	3	0.13
No	41	40	
Escalate Therapy			
Yes	9	3	0.13
No	41	40	
Acute severe colitis			
Yes	5	0	0.06
No	45	43	
Colectomy			
Yes	1	0	$\simeq 1$
No	49	43	
Dysplasia			
Yes	1	1	$\simeq 1$
No	49	42	

The conclusions are the same, most outcomes tend to occur with a higher frequency when there is histological severity, except in the case of dysplasia, where only two individuals present the outcome and are equally distributed, one with histological severity and another without. In acute severe colitis outcome, the p-value is significant. Since all the five individuals that had a acute severe outcome had histological severity, there is an indication that an individual with histological severity is associated with having acute severe colitis.

4.2 Cox Regression models

Before starting to build a model it is necessary to determine, for each patient, if the event occurred, if so, the individual's status is 1, and the time associated with this individual will be a real observed time (the moment it has occurred). If the outcome did not occur, the status will be given by the value 0 and the time associated with the individual will be given as a censorship. The censored time will be calculated from the date of diagnosis until the date of the last record.

For each outcome, the following approach was followed: presentation of an histogram for the observed and censored times, fitting of an univariate model considering the risking factors, and finally, fitting of three multiple models and correspondent Kaplan-Meier curves. All these models include the variable `Extent`, `Age` and `Mayo score` as covariates. To this set of covariates, `Nancy score`, `Involvement beyond endoscopic margin` and `Histological severity` were added, one at a time, resulting in three different models.

In Portugal there has been a considerable consumption of medical resources associated to ulcerative colitis. Between 2010 and 2017, around 3400 people were hospitalised due to this disease, with an average hospital stay of 10 days, with a minimum of 0 days (corresponds to individuals who are hospitalised and do not total 24 hours) and a maximum of 109 days. Therefore, the study of the hospitalisations in this disease is a very relevant issue in order to have a perspective of its occurrence, as well as to help in the monitoring of patients and in clinical and hospital administration decision making. Thus, the analysis for the hospitalisations will be presented. The detailed results, for the other outcomes, can be found in appendix A.

Since this is an exploratory study and a subject under investigation, the significance level to be used is $\alpha = 0.10$, neverthless confidence intervals are presented at 95% confidence.

Simple outcomes

A simple outcome is the outcome that represents the time to a single event. In this case, the response variable will be the time until a certain event.

Hospitalisation

Considering the need of hospitalisation, there were obtained 12 observed times and 81 censored times.

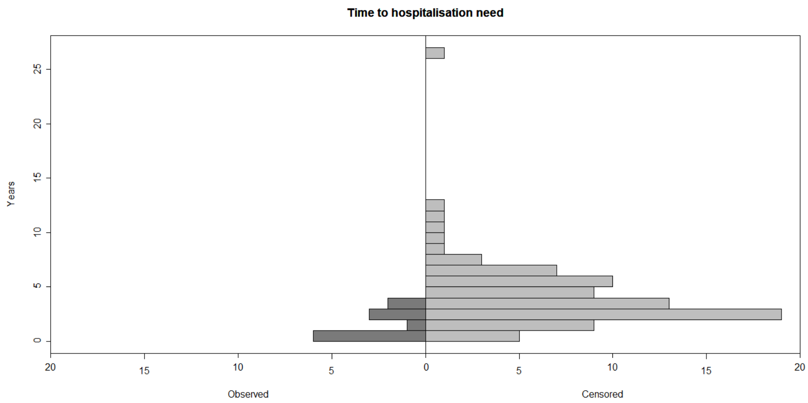


Figure 4.1: Histogram of observed and censored times to hospitalisation.

Figure 4.1 illustrates the distribution of observed and censored times to hospitalisation. It is possible to verify that all patients who have been hospitalised have times between 0 and 4 years (0.02 to 3.31 years). The mean of the observed times to hospitalisation need is 1.34 years. For the censored ones, they have quite superior times. The mean of censored times is 4.33 years, with a minimum of 0.18 years and a maximum of 26.95 years. These times suggest that, although this is a very interesting outcome, there are many individuals who never experience.

In addition to the interest in studying the time until the occurrence of these outcomes, it is of extreme relevance to study which variables provide a higher or lower risk in the occurrence of the outcome. Thus, in Table 4.6 six univariate models are represented, each line represents a different model.

Table 4.6: Cox regression univariate analysis for outcome hospitalisation in patients.

Variable	β	hazard ratio	p-value
Extent - E_2	0.644	1.905	0.265
Age	-0.055	0.947	0.018
Nancy score $\geq$ 3	1.229	3.417	0.240
Mayo score - Moderate and Severe	0.576	1.779	0.388
Involvement beyond margin - Yes	0.726	2.067	0.349
Histological severity	0.996	2.707	0.135

As shown in Table 4.6, the variable that is significantly associated to the need of hospitalisation is `Age`. Older individuals are expected to have later hospitalisations (HR 0.95 (CI 95%: 0.9-0.99), p-value=0.018). By analysing the coefficient of the model, it is possible to verify that age is a protective factor concerning the need of hospitalisation. This conclusion is in agreement with the literature, since it is known that the disease tends to affect mainly younger individuals (**?**).

We now consider `Extent`, `Age`, `Nancy score` and `Mayo score` as covariates. Figure 4.2 shows the Kaplan-Meier curves corresponding to the different combinations of the categorical variables included in the model.

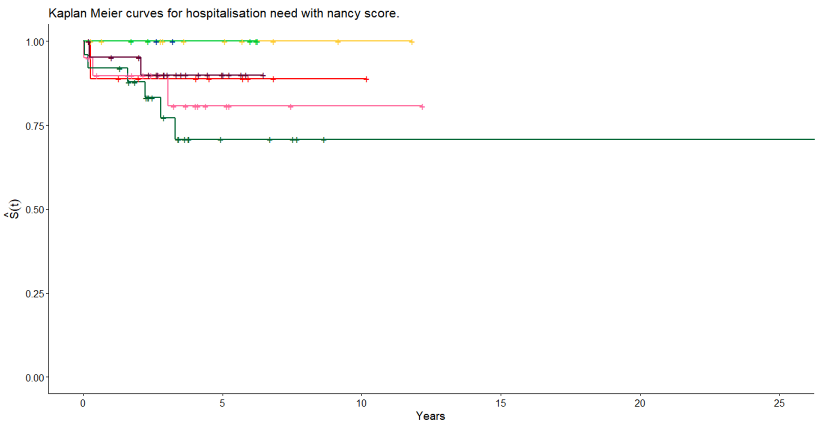


Figure 4.2: Kaplan-Meier estimate of the survival function for time to hospitalisation. ● Extent E_1, Nancy < 3 and Low Mayo Score ● Extent E_1, Nancy < 3 and High Mayo Score ● Extent E_1, Nancy ≥ 3 and Low Mayo Score ● Extent E_1, Nancy ≥ 3 and High Mayo Score ● Extent E_2, Nancy < 3 and Low Mayo Score ● Extent E_2, Nancy < 3 and High Mayo Score ● Extent E_2, Nancy ≥ 3 and Low Mayo Score ● Extent E_2, Nancy ≥ 3 and High Mayo Score

It can be seen that the curve corresponding to strata that has `Extent` E_2, `Nancy score` ≥ 3 and the highest `Mayo score`, is the curve that corresponds to a lower survival. It means that the combination of the most severe categories of the three variables leads to worse survival. The strata that includes the `Extent` E_2, `Nancy score` < 3 and a lower `Mayo score` is not represented in Figure 4.2 since there are no individuals who present this combination.

The fitted model that included the categorical variables and `Age` is presented in Table 4.7.

Table 4.7: Cox regression multiple analysis for outcome hospitalisation, with Nancy score as predictor, in patients.

Variable	β	hazard ratio	p-value
Extent - E_2	0.331	1.393	0.600
Age	-0.055	0.946	0.016
Nancy score ≥ 3	1.281	3.601	0.229
Mayo score - Moderate and Severe	0.162	1.176	0.824

The only variable significantly associated with the hospitalisation need is `Age`. The value 0.946 represents the hazard of hospitalisation needs for two individuals that differ in one year of age, by all the other variables in the model (HR 0.95 (CI 95%: 0.9-0.99), p-value=0.016).

The same was done for the model that includes the variable `Involvement beyond the endoscopic margin`. The Kaplan-Meier curves and the values of this multiple Cox regression are presented in Figure 4.3.

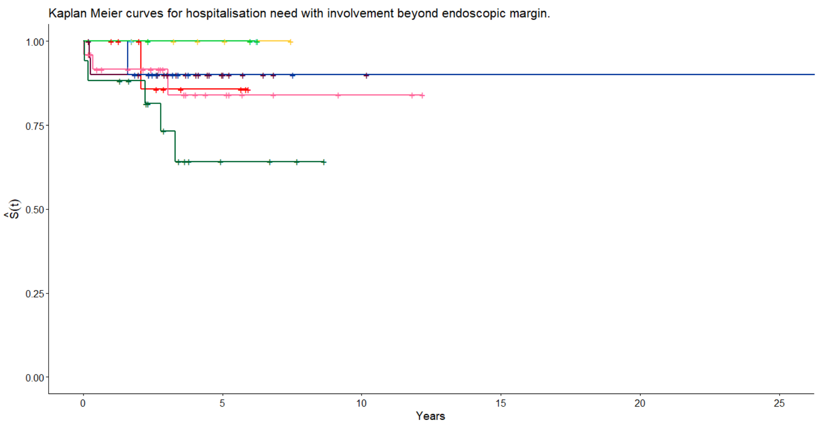


Figure 4.3: Kaplan-Meier estimate of the survival function for time to hospitalisation needs, with the variable involvement beyond endoscopic margin. ● Extent E_1, without Involvement and Low Mayo Score ● Extent E_1, without Involvement and High Mayo Score ● Extent E_1, with Involvement and Low Mayo Score ● Extent E_1, with Involvement and High Mayo Score ● Extent E_2, without Involvement and Low Mayo Score ● Extent E_2, without Involvement and High Mayo Score ● Extent E_2, with Involvement and Low Mayo Score ● Extent E_2, with Involvement and High Mayo Score

The curve with lower survival is the strata that includes the worst categories of each variable, `Extent` E_2, with `Involvement beyond endoscopic margin` and high `Mayo score`.

Table 4.8: Cox regression multiple analysis for outcome hospitalisation, with Involvement beyond margin as predictor, in patients.

Variable	β	hazard ratio	p-value
Extent - E_2	0.594	1.811	0.349
Age	-0.052	0.949	0.020
Involvement beyond margin - Yes	0.895	2.448	0.253
Mayo score - Moderate and Severe	0.321	1.378	0.661

Again, the only statistically significant association is the variable `Age`. Older individuals are expected to have later hospitalisations when compared with younger indivuals (HR 0.95 (CI 95%: 0.91-0.99), p-value=0.02).

The last model is the one where we included `Extent`, `Age`, `Mayo score` and `Histological severity` (`Nancy score` and `Involvement beyond endoscopic margin` combined).

The survival function representation is shown in Figure 4.4.

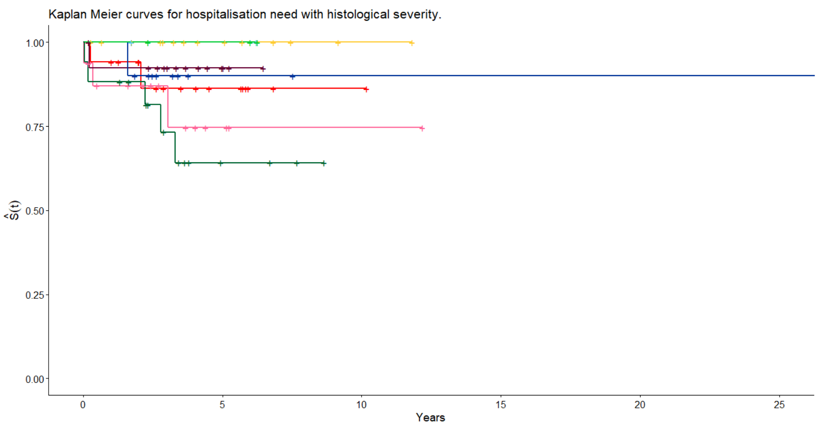


Figure 4.4: Kaplan-Meier estimate of the survival function for time to hospitalisation needs, with the variable histological severity. • Extent E_1, without Histological severity and Low Mayo Score • Extent E_1, without Histological severity and High Mayo Score • Extent E_1, with Histological severity and Low Mayo Score • Extent E_1, with Histological severity and High Mayo Score • Extent E_2, without Histological severity and Low Mayo Score • Extent E_2, without Histological severity and High Mayo Score • Extent E_2, with Histological severity and Low Mayo Score • Extent E_2, with Histological severity and High Mayo Score

Again, the strata representing the combination of the most severe categories leads to lower survival.

The results for the Cox regression analysis is shown in Table 4.9.

Table 4.9: Cox regression multiple analysis for outcome hospitalisation, with Histological severity, in patients.

Variable	β	hazard ratio	p-value
Extent - E_2	0.372	1.450	0.560
Age	-0.052	0.949	0.019
Histological severity	0.975	2.652	0.148
Mayo score - Moderate and Severe	0.317	1.373	0.665

Variable `Age` is strongly associated to the hospitalisation need. For an one year difference in age, the hazard decreases in 0.051, by all the other variables considered in the model (HR 0.95 (CI 95%: 0.91-0.99), p-value=0.019).

It is possible to verify that the variable `Age`, whether it is being tested in a univariate model or in any of the three multiple models, is significant and has a negative coefficient. Thus, regardless of the variables that influence it, of those tested, this variable always causes a decrease in the hazard of an individual being hospitalised.

Use of steroids

For this outcome, there are 20 observed times and 73 censored times.

In the univariate models, the only variable that proved to be significant was `Histological severity` (HR 2.39 (CI 95%: 0.91-6.3), p-value=0.076).

In the multiple model that included the variables `Extent`, `Age`, `Nancy score` and `Mayo score`, there was no variable that proved to be significant. In the model that includes the `Involvement beyond the margin`, this same variable was significant contributing to the steroids use (HR 3.52 (CI 95%: 0.98-12.6), p-value=0.053). As for the model that includes `Histological severity`, this variable is significant and contributes further to the need of steroids use (HR 2.68 (CI 95%: 1.00-7.2), p-value=0.051).

Escalate Therapy

Considering the outcome escalate therapy, there are 12 observed times and 81 censored times, this means that 12 patients experienced the event.

In the univariate models, the variable `Age` is the only variable significantly associated with escalating therapy (HR 0.96 (CI 95%: 0.92-1.00), p-value=0.058).

As for multiple models, considering the model that includes `Extent`, `Age`, `Nancy score` and `Mayo score`, `Age` is the only significantly associated with a patient escalates therapy (HR 0.96 (CI 95%: 0.92-1.00), p-value=0.058). For the multiple model that includes the variable

`Involvement beyond the margin` instead of `Nancy score`, `Age` and `Involvement beyond the margin` proved to be significant in predicting the need of escalate therapy (HR 0.96 (CI 95%: 0.92-1.00), p-value=0.055) and (HR 5.89 (CI 95%: 0.75-46.6), p-value=0.093), respectively.

For the last multiple model, which includes `Histological severity` instead of `Nancy score`, `Age` is negatively associated with the escalation of therapy. Thus, `Age` represents a protective factor in the outcome escalate therapy (HR 0.96 (CI 95%: 0.92-1.00), p-value=0.052).

Acute Severe Colitis, Colectomy and Dysplasia

As for outcomes acute severe colitis, colectomy need and dysplasia, the percentage of censors is very high, with few individuals (less than 6%) having suffered the outcomes, and so it makes no sense to adjust models.

Composed outcomes - Therapeutic needs

During the follow-up, UC patients may have a flare, defined by the reappearance of disease symptoms, such as bloody diarrhoea, abdominal pain and weight loss. After excluding infectious causes, patients usually receive a course of steroids in order to re-induce clinical remission. Nevertheless, steroids are only used in acute flares and are not part of maintenance therapy. When a patient has a severe flare or become steroid-dependent, therapy escalation to other immunosuppressive therapy may be warranted. Thus, a composite outcome was defined that consists of determining the time of the first outcome between two outcomes, in this case steroids use and the need for therapeutic escalation.

There were 21 people who had therapeutic needs and 72 who did not, so these were considered censorships.

Univariate models were again built for each variable, where it was proved that the `Histological severity` is significant (HR 2.6 (CI 95%: 1.00-6.8), p-value=0.05).

When evaluating the multiple model with the variables `Extent`, `Age`, `Nancy score` and `Mayo score`, none of the variables present in the model were significant. In the model with the `Involvement beyond the margin` in place of `Nancy score`, the `Involvement beyond the margin` showed to be significant (HR 3.65 (CI 95%: 1.03-13.0), p-value=0.045). The impact that involvement has on the response variable is of a negative effect for the disease. A person who have `Involvement beyond the margin`, when compared to those who do not, has a higher hazard of therapeutic needs.

Regarding the last model, that has the variable `Histological severity` instead of `Nancy score`, the `Histological severity` is significantly associated with the therapeutic needs (HR 2.93 (CI 95%: 1.10-7.8), p-value=0.031). Individuals with `Histological severity` have higher hazard of needing therapy than those who do not have histological severity.

4.3 Model Diagnostics

In this section the residual analysis will be presented referring to the models implemented in the previous section, considering need of hospitalisation as the outcome. As in the previous section, the analyses for the other models will be presented in Appendix B.

Schoenfeld Residuals

The proportional hazards (PH) assumption can be checked using statistical tests and graphical diagnostics based on the scaled Schoenfeld residuals. For each covariate, the independence between residuals and time it is tested. In principle, the Schoenfeld residuals are independent of time. A plot that shows a non-random pattern against time constitutes evidence of violation of the PH assumption.

First, these patterns were evaluated for the univariate models whose response is the need for hospitalisation. As the conclusions obtained for the univariate models are very similar to the conclusions for the multiple models, only the conclusions of this diagnosis will be presented. For the multiple models, graphic representations, tests and conclusions will be presented.

Since there is no evident pattern, and the p-values are not significant, the proportional hazard assumption is not discarded to each of the univariate models.

The same was made for the multiple models. The results are presented in Figure 4.5.

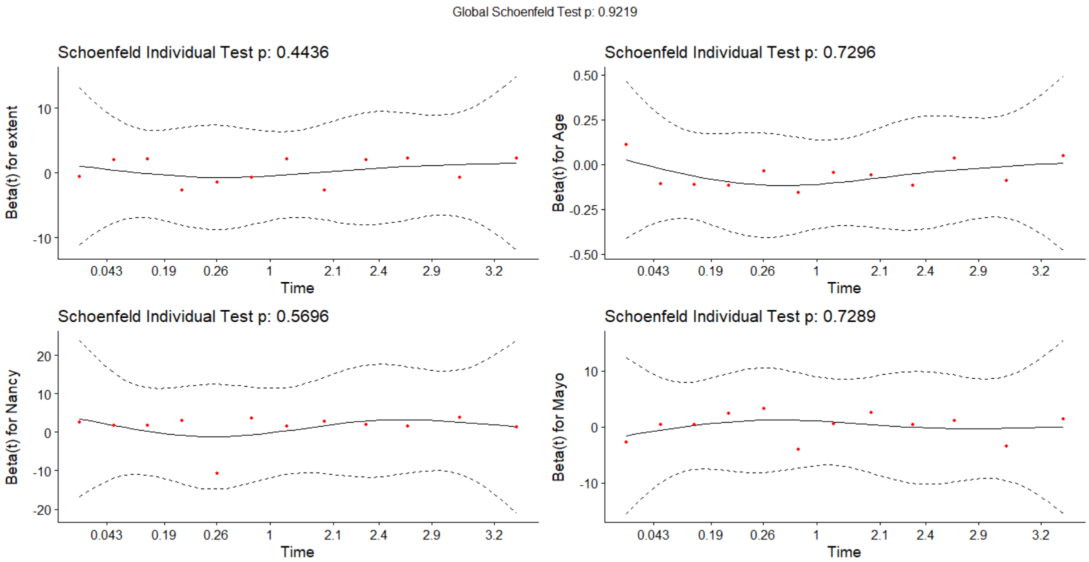


Figure 4.5: Schoenfeld residuals to multiple model with hospitalisation needs, with Nancy score included.

The p-value associated with each variable, as well as the global p-value of the model that includes the four variables (p=0.9219), are none significant. These p-values complemented by the absence of residuals pattern lead us to conclude that the proportional hazard assumption is not violated.

The same analysis was done but considering the model to the variable `Involvement beyond the margin` instead of `Nancy score`.

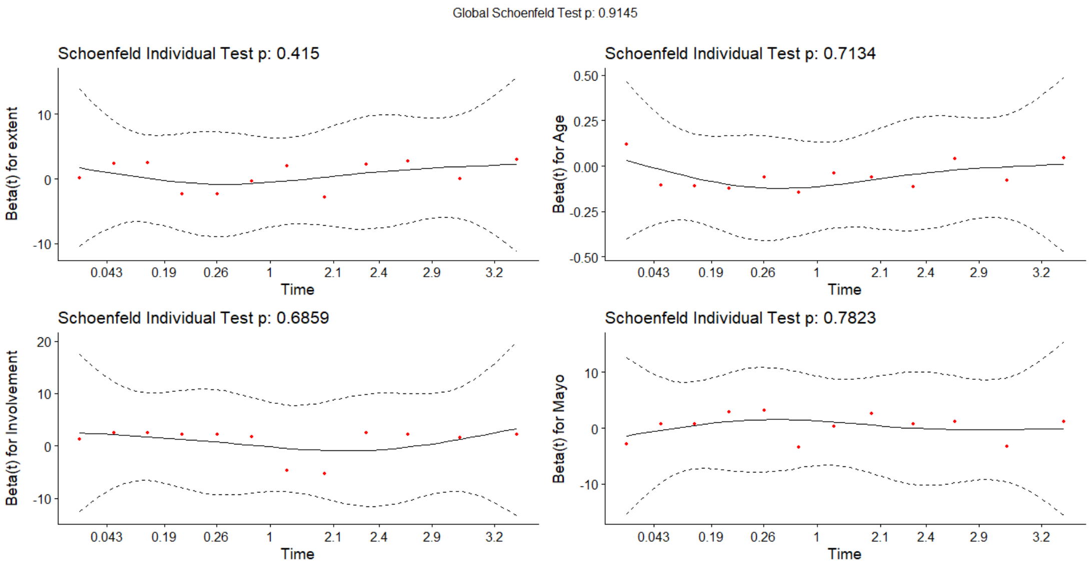


Figure 4.6: Schoenfeld residuals to multiple model with hospitalisation needs, with Involvement beyond margin included.

For this model, the conclusion is the same - the proportional hazards assumption is not violated.

The Schoenfeld residuals for the model that includes the `Histological severity` are presented in Figure 4.7.

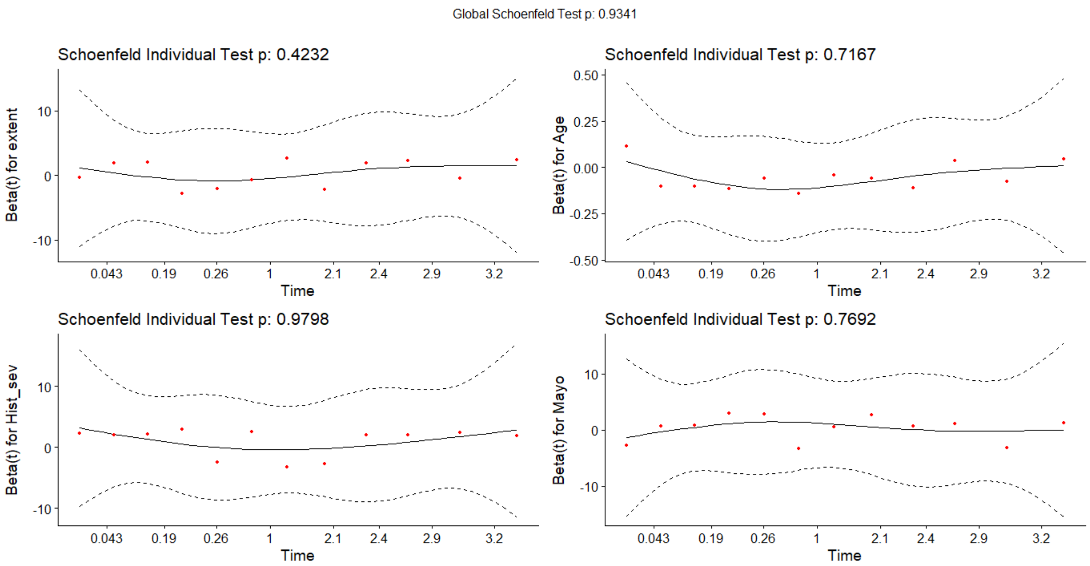


Figure 4.7: Schoenfeld residuals to multiple model with hospitalisation needs, with Histological severity.

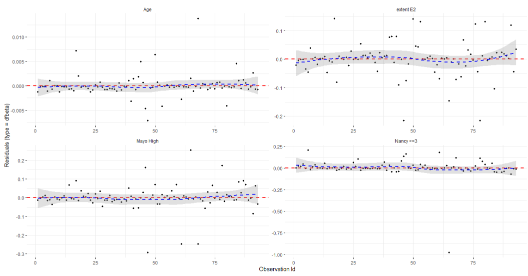
Considering Figure 4.7, the proportional hazards assumption does not seem to be violated.

Dfbeta values and deviance residuals

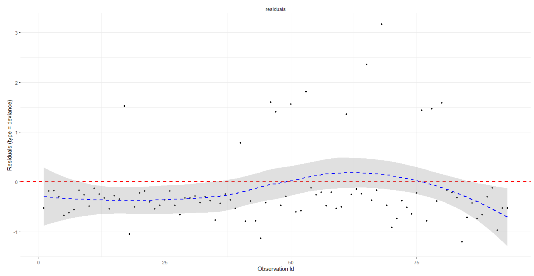
In order to test influential observations or outliers, we can visualize either dfbeta values or deviance residuals.

The dfbeta values plot show, by comparing the magnitudes of the largest dfbeta values to the regression coefficients, that none of the observations is terribly influential individually. It is also possible to check the existence of outliers by visualizing the deviance residuals. The deviance residual is a normalized transform of the martingale residual. These residuals should be roughly symmetrically distributed around zero with a standard deviation of 1. Positive values correspond to individuals that "died too soon" compared to expected survival times. The negative values correspond to individual that "lived too long". Very large or small values are outliers, which are poorly predicted by the model.

The next three Figures 4.8, 4.9 and 4.10 represent the dfbeta values and the deviance residuals for the simple outcome hospitalisation need, including: in the first plot, the `Nancy score`; in the second, the `Involvement beyond the margin` and in the last one `Histological severity`.

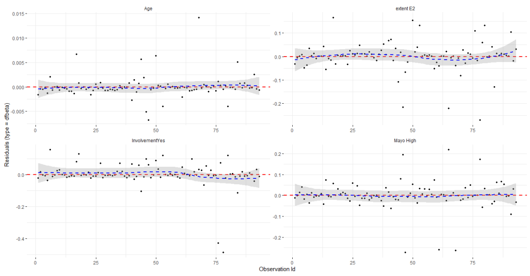


(a) Dfbeta values

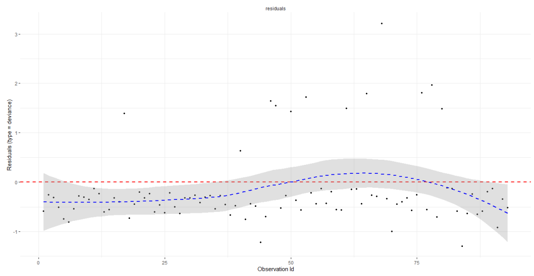


(b) Deviance residuals

Figure 4.8: Dfbeta values and deviance residuals to multiple model to hospitalisation need, with Nancy score.

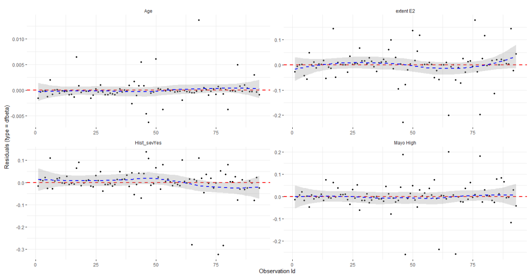


(a) Dfbeta values

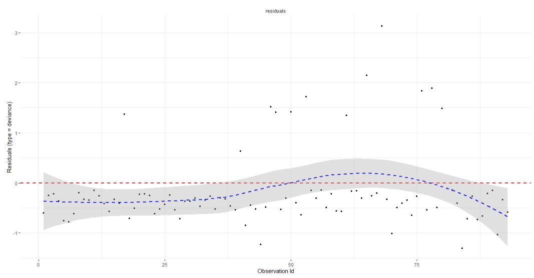


(b) Deviance residuals

Figure 4.9: Dfbeta values and deviance residuals to multiple model to hospitalisation need, with Involvement beyond endoscopic margin.



(a) Dfbeta values



(b) Deviance residuals

Figure 4.10: Dfbeta values and deviance residuals to multiple model to hospitalisation need, with Histological severity.

Since the observations are relatively symmetrically distributed around zero, it can be stated that no influential outliers seem to exist.

The residual analysis of the models in the appendix can also be found in appendix B.

4.4 Cure models

Ulcerative colitis is a chronic disease known to have a long follow-up time, patients can be followed by 27 years. In our retrospective study the average of follow-up time is 4.5 years. However, certain outcomes, such as hospitalisations, are events that do not occur very often (about 85% of censored observations) and also because censored times are substantially higher than the observed times to hospitalisation (median of 3.63 and upper quartile of 5.67 years and median of 0.97 and upper quartile of 2.37 years, respectively). Thus, this leads us to suspect that there are individuals who will never experience this event even if the follow-up time was prolonged indefinitely, motivating the study and application the cure models on our data.

The mixture cure model is a special type of survival model, where it is assumed that the population is a mixture of susceptible individuals, who may experience the event of interest, and cured/non-susceptible individuals, who will never experience the event Cai et al. (2012). In Laska and Meisner (1992), Taylor (1995) and Tamura et al. (2000), individuals followed for more than an *a priori* established amount of time without experiencing the event of interest, are considered cured. In Betensky and Schoenfeld (2001), the occurrence of a certain event (e.g. hospital discharge) distinct from the event of interest (e.g. death from the disease) will enable the identification of the healed individuals.

Undoubtedly, the most common situation is the one where the healed/ non-susceptible individuals can be found among those with the longest censored times.

Since the event of interest in the present study is the occurrence of hospitalisation, we will adopt the terms susceptible/non-susceptible individuals to refer to those who may/will not experience the event.

Figure 4.11 shows a representation of a Kaplan-Meier curve of observed and censored times of ulcerative colitis hospitalisations.

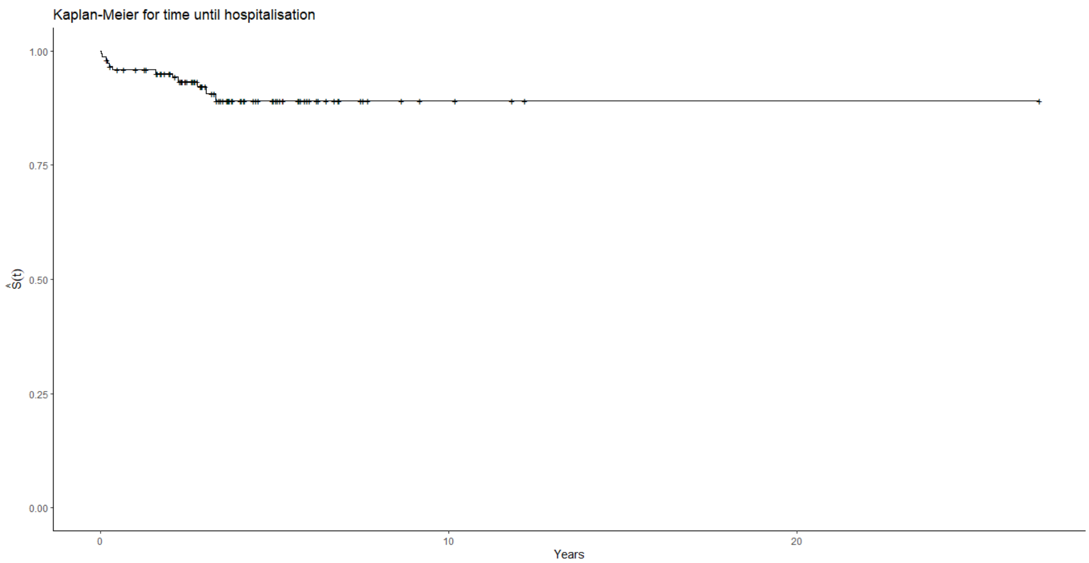


Figure 4.11: Kaplan-Meier for time until hospitalisation.

When comparing times to hospitalisation and censored times, we note that none of the former exceed 3.31 years whereas censored times can be as long as 27 years. Even if we disregard this extreme value, the highest observed time is smaller than the median of the censored times.

This aspect of the data supports the hypothesis that several of these patients will never have the need of hospitalisation. In other words, we would say that they are not susceptible to hospitalisation.

The cure mixture models are composed by two parts, one respecting the (probability of) susceptibility and another concerning the failure time (time up to the occurence of the event).

As far as the susceptibility is concerned, it will be possible to draw conclusions as to which variables will affect the susceptibility of the individuals. As for the components of the failure time, for susceptible individuals (individuals who may suffer from the event hospitalisation need), it is possible to draw conclusions about which variables affect the time to the event.

The mixture cure models were implemented considering the variables `Extent`, `Age`, `Nancy score` and `Mayo score` in both components of the model - susceptibility and failure time distribution. These variables were chosen as explanatory variables since they had already been used in the Cox models, and therefore, it is possible to compare the conclusions of the two methodologies.

Table 4.10 shows the semiparametric proportional hazards mixture cure model with the variables presented above.

Table 4.10: Semiparametric proportional hazards mixture cure model.

Susceptibility component - Logistic regression model:				
Variable	**Estimate**	**Std.Error**	**Z-value**	**p-value**
(Intercept)	-1.154	1.546	-0.746	0.455
Extent	0.490	0.677	0.724	0.469
Age	-0.050	0.040	-1.250	0.211
Nancy score ≥ 3	1.473	0.841	1.751	0.080
Mayo score - Moderate and Severe	0.064	0.779	0.082	0.934

Failure time distribution model:				
Variable	**Estimate**	**Std.Error**	**Z-value**	**p-value**
Extent	-0.365	1.303	-0.280	0.779
Age	-0.052	0.085	-0.608	0.543
Nancy score ≥ 3	-0.664	0.935	-0.711	0.477
Mayo score - Moderate and Severe	0.439	1.461	0.300	0.764

As can be seen from Table 4.10, in the component concerning susceptibility, `Nancy score` is the only relevant variable (p-value=0.08). A higher level of the `Nancy score` favours the susceptibility of an individual when compared to one with a lower Nancy score (< 3) (OR=4.36; CI 90% (1.09,17.4)).

When it comes to the failure time component, no variables were significant, meaning that no evidence was found to claim that any of the variables influence significantly the time to the occurrence of an hospitalisation among the susceptible individuals.

Although non particular significant results were found, it is interesting to explore some of the information that can be extracted from these models, in particular concerning the susceptibility component. It is possible to estimate the probability that an individual will never experience hospitalisation, given his/her profile. In Figure 4.12 the probability of being non-susceptible (and hence never need hospitalisation) can be seen as a function of `Age`, for the 8 combinations of `Extent`, `Nancy score` and `Mayo score`.

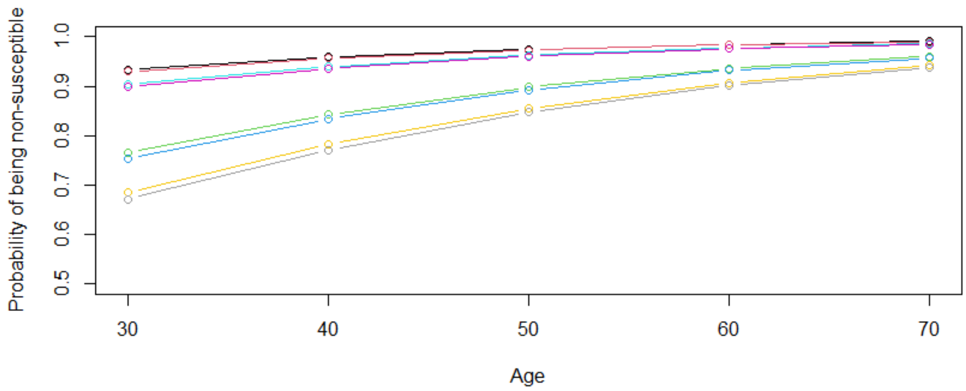


Figure 4.12: Probability of being non-susceptible by Age. • Extent E_1, Nancy < 3 and Low Mayo Score • Extent E_1, Nancy < 3 and High Mayo Score • Extent E_1, Nancy ≥ 3 and Low Mayo Score • Extent E_1, Nancy ≥ 3 and High Mayo Score • Extent E_2, Nancy < 3 and Low Mayo Score • Extent E_2, Nancy < 3 and High Mayo Score • Extent E_2, Nancy ≥ 3 and Low Mayo Score • Extent E_2, Nancy ≥ 3 and High Mayo Score

The mixture cure models seem to be an interesting choice in the presence of a large proportion of censored observations even after a long follow-up. Besides producing an estimate of the probability of "cure" given the patient's profiles, it is expected to provide a more realistic estimate of the survival function for the fraction of the population who will suffer the event.

When observing the failure time distribution part, it can be seen, in relation to the Cox model, that the effect of the `Extent` and `Nancy score` variables has changed. In this component of the cure mixture model, "weights" are added in the calculation of the individual's survival function, which may have caused a change in the parameters. Even so, despite the change in the effects on these variables, this was not the biggest change. The variable `Age`, which in the Cox model contributed significantly, in the cure model is not significant.

Since only the `Nancy score` proved to be significant, another model was estimated with this variable only, both in susceptibility component and the failure time part.

Table 4.11: Semiparametric proportional hazards mixture cure model with variable Nancy score.

Susceptibility component - Logistic regression model:				
Variable	**Estimate**	**Std.Error**	**Z-value**	**p-value**
(Intercept)	-2.913	0.560	-5.205	$<< 0.001$
Nancy score ≥ 3	1.470	0.650	2.262	0.024
Failure time distribution model:				
Variable	**Estimate**	**Std.Error**	**Z-value**	**p-value**
Nancy score ≥ 3	0.924	0.655	-1.411	0.158

Using Table 4.11, we can conclude about the influence that the `Nancy score` variable has on the susceptibility component (p-value=0.024).

The cure rate can be calculated based on the results from *Cure probability model* part, since it is the part that deals with the estimation of the proportion of the cured and uncured in the model. In this case, with the model shown in Table 4.11, we can calculate two cure rates, for individuals with `Nancy score` < 3 and for individuals with `Nancy score` ≥ 3.

The estimated cure rates are of 94.8% and 80.9% for patients with `Nancy score` < 3 and ≥ 3 respectively, indicating that a very high percentage of the patients will never experience hospitalisation. Regarding the failure time part of the model, although `Nancy score` presented a smaller p-value when compared to the previous model, still it was not significant. However, the coefficient makes sense as it points towards a higher hazard for hospitalisation for the group with higher values of the `Nancy score`.

Due to the complexity of the estimating equation in the EM algorithm, the standard errors of estimated parameters are not directly available Cai et al. (2012). Bootstrap sampling is used to overcome this difficulty. In the R package by Cai et al. (2012) a pre-determined number of bootstrap samples are drawn, each providing set of parameters estimates. Bootstrap confidence intervals and then computed. In our case, possibly due to the small number of non-censored observations, it was not possible to use this functionality automatically since the estimation procedure would not converge. The solution found consisted in reproducing the same approach in a more controlled manner, keeping the results from the runs that converged and discarding the cases where no convergence was met.

The estimation algorithm converged 105 out of 300 times. This number of estimates of the model parameters were then used to calculate bootstrap confidence intervals.

Chapter 5

Discussion and Conclusions

This work intended to contribute to the understanding of the prognostic value, if any, of the presence of histological inflammation above the endoscopic margin at the time of diagnosis in medium and long-term outcomes. In order to meet the main objective of the work, information from patients, such as socio-demographic and clinical variables, was gathered.

Age at diagnosis was also collected, as previous studies concluded that the disease is more prevalent among younger individuals. Clinical information such as the extent of the disease, the severity of the disease at the time of diagnosis and other histological variables that assess the severity of the disease at the time of follow-up were also collected.

In a first phase, in order to respond to the objective of the study, and taking into account that we had access to these variables and to the times until the outcomes occur, it was decided to use Cox regression models. Cox models were used since proportional hazards assumption was not rejected and this type of models are popular amongst medical researchers, especially concerning the parameter interpretation. Moreover, according to Nardi and Schemper (2003), a good discrimination among parametric models require censoring not to exceed 40-50 per cent. Our data is of moderate size and about 85% of the cases correspond to censored times. Nevertheless, for comparison purposes, we consider fitting parametric models in future work since they can perform better than the Cox model if the conditions they require are met.

Special attention was given to the study of the hospitalisations outcome, since in Portugal there has been a considerable consumption of medical resources associated to UC, making the study of hospitalisations in this disease is a very relevant issue in order to have a perspective of its occurrence, as well as to help in the monitoring of patients and in clinical and hospital administration decision making.

For the study of this outcome, all the 93 patients were considered, 12 of which had been hospitalised at some time and the remaining were considered censored. Variables age at diagnosis, extent of the disease at diagnosis, endoscopic severity at diagnosis, histological severity and involvement beyond the endoscopic margin were included as covariates. The `Histological severity` results from the combination of the `Nancy score` and `Involvement beyond the margin`. This variable has two levels, one of the levels corresponds to the most severe categories of the two variables (`Nancy score` ≥ 3 and with `Involvement beyond the margin`) and the other level with the remaining categories. After the study of the Cox models, and given the high percentage of censorships, the hypothesis that individuals who had not suffered the event until then would probably not suffer it even if the study was prolonged was raised. Thus, with interest in studying whether the censored individuals might be non susceptible to hospitalisations or whether they would suffer the event later, the idea of studying the cure models emerged. The 93 individuals were assessed on the severity of the disease, at the time of diagnosis and at the time of follow-up, where a high percentage (%) of severely ill individuals was found. Next, it was decided to assess whether these variables had an influence on the occurrence of the studied outcomes, i.e., whether a higher severity would be associated with the presence of the outcome. Although this relationship is not significant in any of the outcomes, a higher severity has no significant influence on the occurrence of the outcomes, these outcomes tend to occur more in individuals with higher severity of disease, either at the time of diagnosis or at the time of follow-up.

In order to verify how the combination of the different variables influences the survival of individuals, strata were formed with the different combinations of the variables under study in order to include all possible profiles of the individuals under study. When observing these representations, it was concluded that although not very evident, as they do not differ greatly from each other, individuals who are in the stratum that combines the most severe categories of variables, have a lower survival. In other words, an individual with the most severe categories, `Extent` E2, a higher `Nancy score` and `Mayo score` have a lower survival.

In order to support this conviction and verify which variables could be associated with a worse prognosis, the Cox regression models were constructed. From these models, we concluded that the only variable that was found to be statistically associated with the risk of an individual being hospitalised is `Age`. As already concluded in other studies, `Age` acts as a protective effect on the risk of an individual needing to be hospitalised. If it is a protective factor, the hazard of an individual being hospitalised decreases as age increases. Yonger age at diagnosis has been consistently a risk factor for shorter time to clinical relapse. If there is a unitary increase in `Age`, the hazard of an individual being hospitalised decreases by about 0.051.

With respect to the other outcomes under analysis, such as the use of steroids, we both associate higher `Histological severity` and `Involvement beyond the margin` to an increase of the hazard for corticoid use. As for the outcome of escalate therapy, where `Age` and `Involvement beyond the endoscopic margin` variables contribute to the increased risk of a patient to escalate therapy. When studying the behaviour of the times until the occurrence of hospitalisations, it was found that, from the Kaplan-Meier curve, the existence of observed times occurred in the first years of the study, and the remaining years were marked solely by censorship times. This situation raised the hypothesis that the need for a patient to be hospitalised is not a very frequent event and that perhaps, even if the study was prolonged for more years, they would never suffer the event. Some of the censored individuals are only individuals who have not suffered the event until then, while others will never suffer, the last being considered non-susceptible. One of the main interests was to calculate what percentage of the population was non-susceptible, as well as to verify which variables could influence this choice.

In the first adjusted model, which included the `Extent`, `Age`, `Nancy score` and `Mayo score` variables, the only variable that was significantly associated with the susceptibility is the `Nancy score`. It was concluded that a higher `Nancy score` level favored an individual's susceptibility. As for the failure time part, no variable was considered significant. In order to calculate the cure rates for two individuals with different profiles in the `Nancy score` (< 3 and ≥ 3), it was built an univariate model that includes only this variable. This model is important since this variable is the only one that is significantly associated with the susceptibility. The estimated cure rates are of 94.8% and 80.9% for patients with `Nancy score` < 3 and ≥ 3 respectively, indicating that a very high percentage of the patients will never experience hospitalisation. Regarding the failure time part of the model, although `Nancy score` presented a smaller p-value when compared to the previous model, still it was not significant.

This work allowed us to study something that had never been studied before because it is the first study to assess the prognostic value of histological features at the time of diagnosis in treatment-naïve E1 and E2 UC patients. It allowed to include patients from several hospital centres and to use a validated score for this disease. Much can be improved as future work if one of the major limitations of this study, the sample size, is overcome. Due to the fact that the sample is small and the events in question are infrequent, the results obtained throughout the study were not statistically significant and therefore could not be used to draw conclusions for clinical purposes in order to advise the medical teams. However, since this is an exploratory study and as it is the first study in this area, the conclusions that are drawn, for example, which variables are good predictors for the occurrence of outcomes, can be used as a good indicator and starting point for future research. Although the follow-up time for this study is quite long, only few individuals tend to be hospitalised, raising the hypothesis that many may never suffer the event. This hypothesis was studied using the mixture cure models.

In order to overcome the sample size problem, in the future, more patients from other centers could be included, in order to use other methodologies, if the assumptions so allowed.

It would also include new variables that could be interesting for the outcomes of interest, as well as including variables at different times of follow-up. The last one would allow patients to have several assessments of the same variable at different times in the study, which could allow the study of recurrent events.

Bibliography

Aalen, O. O., Ø. Borgan, and H. K. Gjessing
2008. *Survival and Event History Analysis*.

Abreu, A. M.
2005. *Modelos de Sobrevivência para Populações com Indivíduos Imunes*. PhD thesis, University of Madeira.

Basson, M. D.
2015. Ulcerative Colitis. P. 2019.

Betensky, R. A. and D. A. Schoenfeld
2001. Nonparametric Estimation in a Cure Model with Random Cure Times. (March):282–286.

Bryant, R. V., D. C. Burger, J. Delo, A. J. Walsh, S. Thomas, A. Von Herbay, O. C. Buchel, L. White, O. Brain, S. Keshav, B. F. Warren, and S. P. Travis
2016. Beyond endoscopic mucosal healing in UC: Histological remission better predicts corticosteroid use and hospitalisation over 6 years of follow-up. *Gut*, 65(3):408–414.

Cai, C., Y. Zou, Y. Peng, and J. Zhang
2012. smcure : An R-package for estimating semiparametric mixture cure models. *Computer Methods and Programs in Biomedicine*, 108(3):1255–1260.

Chen, Z.
2000. A new two-parameter lifetime distribution with bathtub shape or increasing failure rate function. *Statistics and Probability Letters*, 49(2):155–161.

Collett, D.
2003. *Modelling Survival Data in Medical Research*, 2nd edition.

Colosimo, E. A. and S. R. Giolo
2006. *Análise de Sobrevivência Aplicada*, 1st edition.

Cox, D. and D. Oakes
1984. *Analysis of Survival Data*.

Hosmer, D. W., S. Lemeshow, and S. May
2008. *Applied Survival Analysis - Regression Modeling of Time to Event Data*, 2nd edition.

Kleinbaum, D. G. and M. Klein
2012. *Survival Analysis - A Self-Learning Text*, 3rd edition.

Laska, E. M. and M. J. Meisner
1992. Nonparametric Estimation and Testing in a Cure Model. 48(4):1223–1234.

Li, J. and S. Ma
2013. *Survival Analysis in Medicine and Genetics*.

Liu, X.
2012. *Survival Analysis Models and Applications*.

Loftus, E. V.
2004. Clinical Epidemiology of Inflammatory Bowel Disease : Incidence , Prevalence , and Environmental Influences. Pp. 1504–1517.

Moore, D. F.
2016. *Applied Survival Analysis Using R*.

Nardi, A. and M. Schemper
2003. Comparing Cox and parametric models in clinical studies. *Statistics in Medicine*, 22(23):3597–3610.

Ordas, I., L. Eckmann, M. Talamini, D. C. Baumgart, and W. J. Sandborn
2012. Ulcerative colitis. *Psychosomatics*, 24(6):515–519.

Rassen, J. A., S. Schneeweiss, and M. Alexander
2011. 596. Safety of Adjuvanted Pandemic Influenza Vaccines. (Ci).

Rocha, C. and A. L. Papoila
2009. *Análise de Sobrevivência*. Sociedade Portuguesa de Estatística.

Tamura, R. N., D. E. Faries, and J. Feng
2000. Comparing time to onset of response in antidepressant clinical trials using the cure model and the Cramer-von Mises test. (October 1999):2169–2184.

Taylor, J. M. G.
1995. Semi-parametric Estimation in Failure Time Mixture Models. 51(3):899–907.

Appendix

A Cox regression models

Steroids Use

Considering the model that predicts the need of steroids use, it was obtained 20 observed times and 73 censored times.

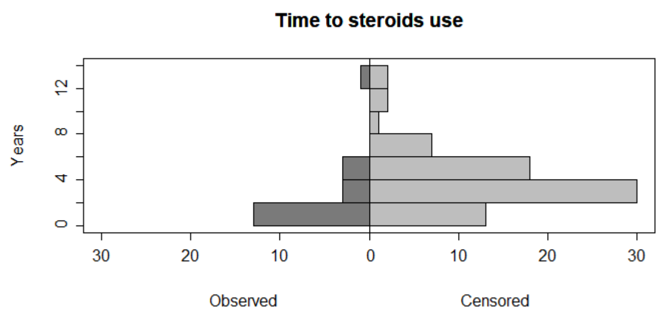


Figure 1: Histogram of observed and censored times to steroids.

Figure 1 represents the observed and censored times for the steroids use. For the individuals that needed a steroids course, the times were between 0.04 and 12.18 years. The mean observed time to the steroids use is 2.17 years. As for the censored times, the times are distributed between 0.18 and 12.82 years. The mean time of censored times is 4.07 years. These times do not seem to differ much in terms of range, however in average terms the times are higher in censored individuals.

In the Table 1, it will be represented six univariate models, each line represents a different model.

Table 1: Cox regression univariate analysis for outcome steroids in patients.

Variable	β	hazard ratio	p-value
Extent - E_2	0.463	1.588	0.304
Age	-0.011	0.989	0.461
Nancy score $\geq$ 3	0.609	1.838	0.335
Mayo score - Moderate and Severe	0.762	2.143	0.144
Involvement beyond margin - Yes	0.959	2.610	0.126
Histological severity	0.873	2.393	0.076

As shown in Table 1, the variable histological severity is the only significantly associated with the steroids use. A patient that has `Histological severity` (`Nancy score` ≥ 3 and with `Involvement beyond endoscopic margin`) has a chance of steroids use 1.393 times bigger than a patient without `Histological severity`.

After analyzing the univariate models, we also evaluated the multiple models. Considering `Extent`, `Age`, `Nancy score` and `Mayo score` as covariates, the estimated survival function for the time until the use of steroids is presented in Figure 2.

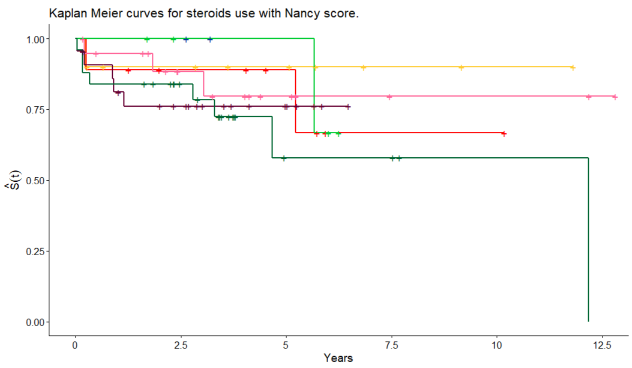


Figure 2: Kaplan-Meier estimate of the survival function for time to steroids use. ● Extent E_1, Nancy < 3 and Low Mayo Score ● Extent E_1, Nancy < 3 and High Mayo Score ● Extent E_1, Nancy ≥ 3 and Low Mayo Score ● Extent E_1, Nancy ≥ 3 and High Mayo Score ● Extent E_2, Nancy < 3 and Low Mayo Score ● Extent E_2, Nancy < 3 and High Mayo Score ● Extent E_2, Nancy ≥ 3 and Low Mayo Score ● Extent E_2, Nancy ≥ 3 and High Mayo Score

As we can see in the Figure 2, the curve corresponding to strata that has `Extent` E_2, `Nancy score` ≥ 3 and the highest `Mayo score`, is the curve that corresponds to a lower survival. This means that the combination of the most severe categories of the three variables all combined leads to worse survival.

In Table 2 is presented a multiple model that includes the categorical variables `Extent`, `Nancy score` and `Mayo score` and `Age`.

Table 2: Cox regression multiple analysis for outcome steroids, with Nancy score as predictor, in patients.

Variable	β	hazard ratio	p-value
Extent - E_2	0.133	1.143	0.786
Age	-0.010	0.990	0.478
Nancy score $\geq$ 3	0.545	1.725	0.410
Mayo score - Moderate and Severe	0.653	1.920	0.233

In Table 2, there are no significant variables to predict the steroids use. The combined effect of the variables is not significant.

The same was made for the model that includes the variable `Involvement beyond the endoscopic margin` instead of `Nancy score`.

Then, the Kaplan-Meier curves for the time until the use of steroids with the effect of `Involvement beyond the endoscopic margin` is shown in Figure 2.

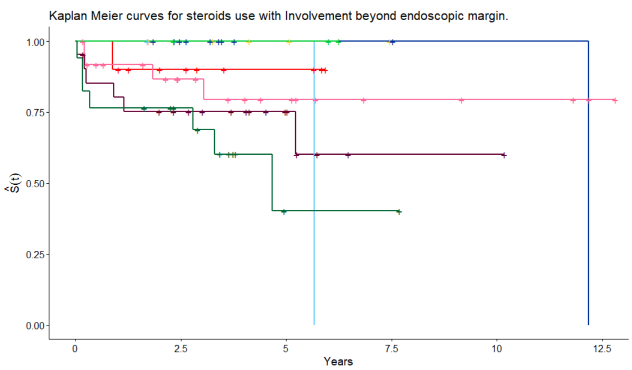


Figure 3: Kaplan-Meier estimate of the survival function for time to use of steroids, with the variable Involvement beyond margin. • Extent E_1, without Involvement and Low Mayo Score • Extent E_1, without Involvement and High Mayo Score • Extent E_1, with Involvement and Low Mayo Score • Extent E_1, with Involvement and High Mayo Score • Extent E_2, without Involvement and Low Mayo Score • Extent E_2, without Involvement and High Mayo Score • Extent E_2, with Involvement and Low Mayo Score • Extent E_2, with Involvement and High Mayo Score

In the Figure 3, the curve corresponding to strata that has `Extent` E_2, with `Involvement beyond the margin` and the highest `Mayo score`, is the curve that corresponds to a lower survival.

The multiple model that was built was the model with the `Involvement beyond the endoscopic margin` in place of `Nancy score`. This model will be presented in Table 3.

Table 3: Cox regression multiple analysis for outcome steroids, with Involvement beyond margin as predictor, in patients.

Variable	β	hazard ratio	p-value
Extent - E_2	0.412	1.509	0.389
Age	-0.011	0.990	0.465
Involvement beyond margin - Yes	1.258	3.518	0.053
Mayo score - Moderate and Severe	0.832	2.298	0.139

It is possible to see, in Table 3, that the variable `Involvement beyond the endoscopic margin` is significantly associated to the need of steroids use. For a person that has involvement beyond the endoscopic margin, the hazard of steroids use increases 2.518 comparing to an individual without involvement, by all the other variables in the model.

The last model is the one where we included `Extent`, `Age`, `Histological severity`.

The survival function to this multiple model, that includes only the categorical variables is shown in Figure 4.

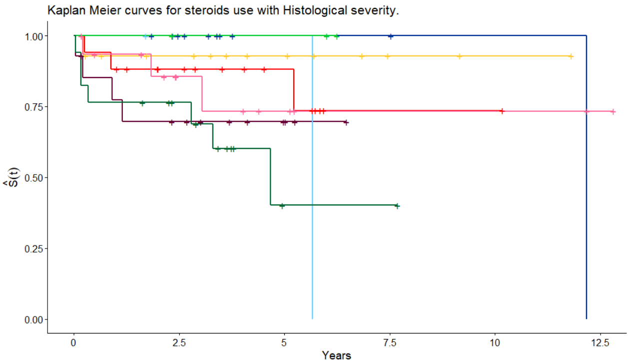


Figure 4: Kaplan-Meier estimate of the survival function for time to use of steroids, with the variable Histological severity. • Extent E_1, without Histological severity and Low Mayo Score • Extent E_1, without Histological severity and High Mayo Score • Extent E_1, with Histological severity and Low Mayo Score • Extent E_1, with Histological severity and High Mayo Score • Extent E_2, without Histological severity and Low Mayo Score • Extent E_2, without Histological severity and High Mayo Score • Extent E_2, with Histological severity and Low Mayo Score • Extent E_2, with Histological severity and High Mayo Score

In Figure 4 leads us to conclude that the worst categories of the categorical variables all combined leads to a worse survival.

Table 4 represents the Cox regression to the model that includes the categorical variables presented above and `Age`.

Table 4: Cox regression multiple analysis for outcome steroids, with Histological severity, in patients.

Variable	β	hazard ratio	p-value
Extent - E_2	0.134	1.144	0.778
Age	-0.012	0.988	0.390
Histological severity	0.985	2.678	0.051
Mayo score - Moderate and Severe	0.812	2.253	0.148

`Histological severity` is the only variable significantly associated with the steroids use.

A patient with histological severity has a hazard of steroids use 1.678 higher compared with a patient without it, by all the other variables.

Escalate Therapy

Considering the outcome escalate therapy, we obtained 12 observed times and 81 censored times. The representation of this times is shown in Figure 5.

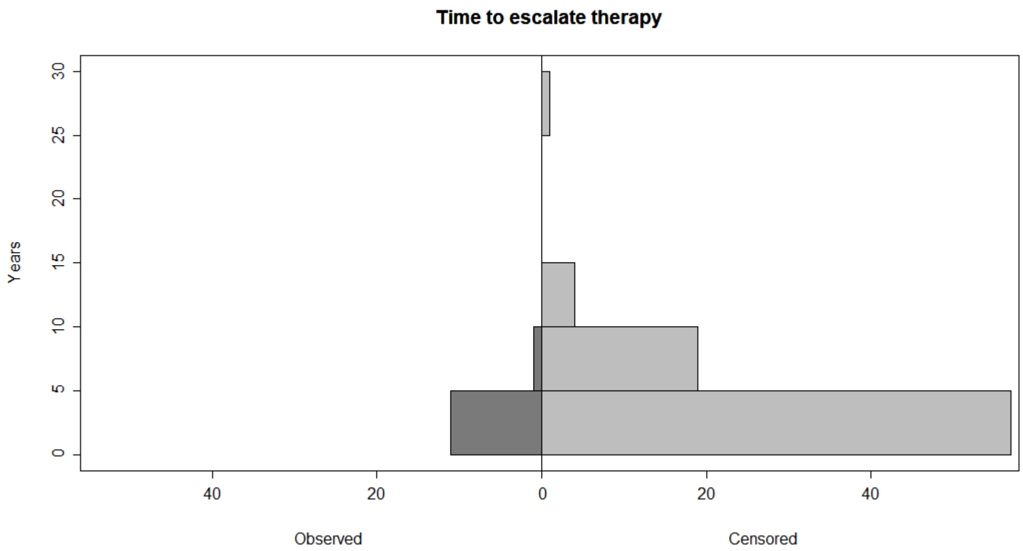


Figure 5: Histogram of observed and censored times to escalate therapy.

The observed times of the individuals who escalate therapy vary between 0.08 and 5.96 years, having a mean time of 2.05 years. As for the censored times, these vary between 0.18 and 26.95 years, and the mean time is 4.33 years. Although there is a censored observation with a very high time, even if we did not consider it, the censored times were going to be higher than the observed times, as suggested by the histogram represented.

In Table 5 there are represented six univariate models, each line represent a different model whose response variable is the escalate therapy needs.

Table 5: Cox regression univariate analysis for outcome escalate therapy in patients.

Variable	β	hazard ratio	p-value
Extent - E_2	0.627	1.871	0.279
Age	-0.040	0.961	0.058

Variable	β	hazard ratio	p-value
Nancy score $\geq$ 3	0.493	1.637	0.526
Mayo score - Moderate and Severe	0.300	1.351	0.626
Involvement beyond margin - Yes	1.542	4.675	0.140
Histological severity	1.070	2.915	0.109

As we can see from Table 5, `Age` is the only variable that is significant, with older individuals being expected to escalate therapy later.

We now consider `Extent`, `Nancy score` and `Mayo score` as covariates. The estimated survival function for the time until the need of escalate therapy with the effect this covariates is shown in Figure 6.

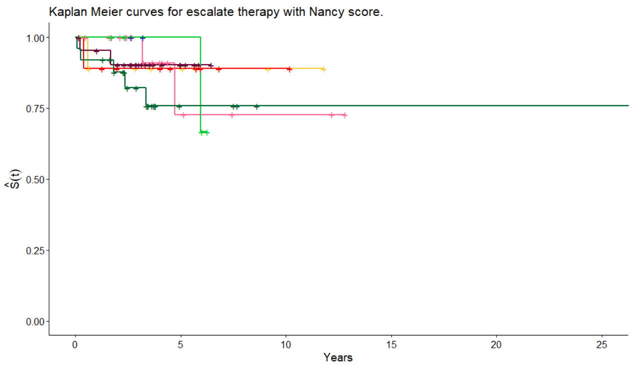


Figure 6: Kaplan-Meier estimate of the survival function for time to escalate therapy. • Extent E_1, Nancy < 3 and Low Mayo Score • Extent E_1, Nancy < 3 and High Mayo Score • Extent E_1, Nancy ≥ 3 and Low Mayo Score • Extent E_1, Nancy ≥ 3 and High Mayo Score • Extent E_2, Nancy < 3 and Low Mayo Score • Extent E_2, Nancy < 3 and High Mayo Score • Extent E_2, Nancy ≥ 3 and Low Mayo Score • Extent E_2, Nancy ≥ 3 and High Mayo Score

Table 6 represents the Cox model for the covariates presented above and `Age`.

Table 6: Cox regression multiple analysis for outcome escalate therapy, with Nancy score as predictor, in patients.

Variable	β	hazard ratio	p-value
Extent - E_2	0.478	1.613	0.467
Age	-0.040	0.961	0.058
Nancy score $\geq$ 3	0.432	1.541	0.598
Mayo score - Moderate and Severe	-0.015	0.985	0.982

According to Table 6, `Age` is the only variable that is significant. `Age` acts as a protective factor, meaning that older individuals are expected to haver later courses of steroids.

The other model is the one where there is included `Involvement beyond the margin` instead of `Nancy score`. The Kaplan-Meier curve to this model is presented in Figure 7.

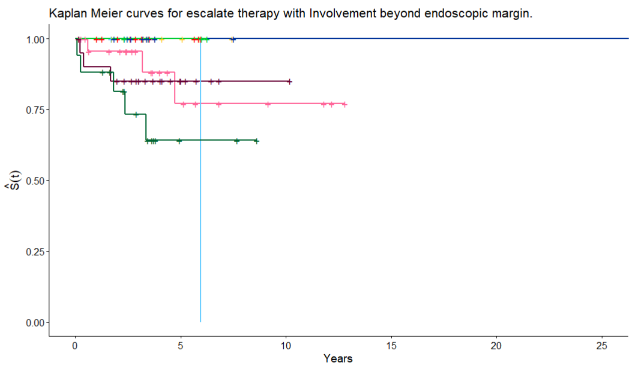


Figure 7: Kaplan-Meier estimate of the survival function for time to escalate therapy, to Involvement beyond the endoscopic margin. ● Extent E_1, without Involvement and Low Mayo Score ● Extent E_1, without Involvement and High Mayo Score ● Extent E_1, with Involvement and Low Mayo Score ● Extent E_1, with Involvement and High Mayo Score ● Extent E_2, without Involvement and Low Mayo Score ● Extent E_2, without Involvement and High Mayo Score ● Extent E_2, with Involvement and Low Mayo Score ● Extent E_2, with Involvement and High Mayo Score

Again, the strata representing the combination of the most severe categories leads to lower survival.

The multiple model of these categorical variables and `Age` will be demonstrated in Table 7.

Table 7: Cox regression multiple analysis for outcome escalate therapy, with Involvement beyond margin as predictor, in patients.

Variable	β	hazard ratio	p-value
Extent - E_2	0.755	2.129	0.230
Age	-0.041	0.960	0.055
Involvement beyond margin - Yes	1.774	5.892	0.093
Mayo score - Moderate and Severe	0.118	1.125	0.862

The variable that, in this multiple model, is significant is `Involvement beyond the endoscopic margin`. For a patient with involvement beyond the margin, the hazard of escalate therapy is 4.892 higher comparing to a patient without involvement beyond endoscopic margin, considering the other variables in the model.

The last model produced is the model where is included `Extent`, `Age`, `Mayo score` and `Histological severity`.

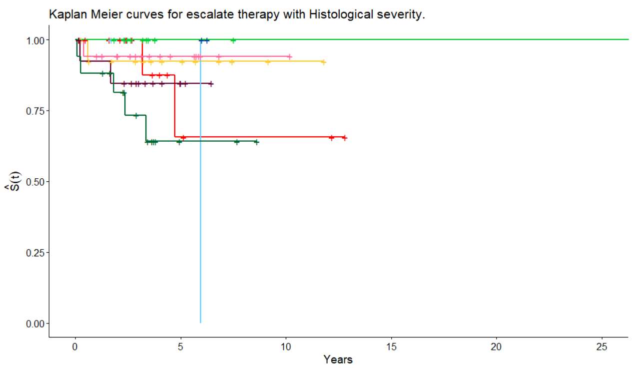


Figure 8: Kaplan-Meier estimate of the survival function for time to escalate therapy, to Histological severity. • Extent E_1, without Histological severity and Low Mayo Score • Extent E_1, without Histological severity and High Mayo Score • Extent E_1, with Histological severity and Low Mayo Score • Extent E_1, with Histological severity and High Mayo Score • Extent E_2, without Histological severity and Low Mayo Score • Extent E_2, without Histological severity and High Mayo Score • Extent E_2, with Histological severity and Low Mayo Score • Extent E_2, with Histological severity and High Mayo Score

The strata representing the combination of the most severe categories leads to lower survival.

The Cox regression to the multiple model where the response variable is the escalate therapy is presented below.

Table 8: Cox regression multiple analysis for outcome escalate therapy, with Histological severity, in patients.

Variable	β	hazard ratio	p-value
Extent - E_2	0.408	1.504	0.522
Age	-0.040	0.961	0.052
Histological severity	1.101	3.009	0.106
Mayo score - Moderate and Severe	0.141	0.683	0.837

`Age` is the only variable significantly associated with the need of escalate therapy. Older individuals are expected to have later episodes of escalating therapy.

Acute severe colitis

Considering the episode of acute severe colitis, there were obtained 5 observed times and 88 censored times. Since there is only 5 observed times, the histogram in Figure 9 has that aspect.

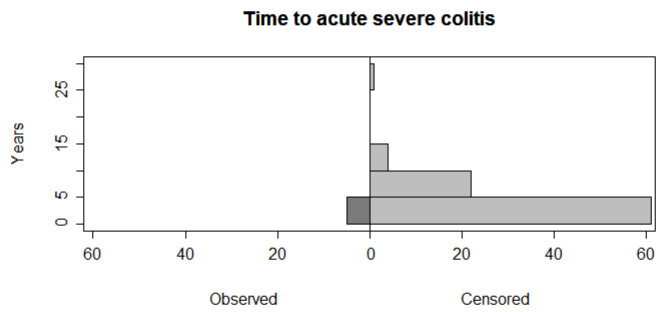


Figure 9: Histogram of observed and censored times to acute severe colitis.

The five patients with this outcome, acute severe colitis, have times between 0 and 5 years (0.05 to 3.31 years), and a mean time of 1.90 years. When it comes to censored times, they already have a bigger range (0.18 to 26.95 years) and a mean of 4.35 years.

Since only 5.4 % had an acute severe colitis, it makes no sense to build models because they will not give great information. However, it will be presented the Kaplan-Meier curves to the multiple models.

The estimated survival function for the time until a patient has an episode of acute severe colitis with the effect of `Extent`, `Nancy score` and `Mayo score` is shown in Figure 10.

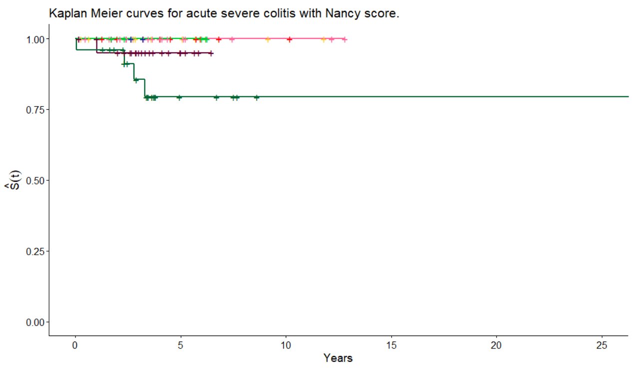


Figure 10: Kaplan-Meier estimate of the survival function for time to acute severe colitis. • Extent E_1, Nancy < 3 and Low Mayo Score • Extent E_1, Nancy < 3 and High Mayo Score • Extent E_1, Nancy ≥ 3 and Low Mayo Score • Extent E_1, Nancy ≥ 3 and High Mayo Score • Extent E_2, Nancy < 3 and Low Mayo Score • Extent E_2, Nancy < 3 and High Mayo Score • Extent E_2, Nancy ≥ 3 and Low Mayo Score • Extent E_2, Nancy ≥ 3 and High Mayo Score

The curve with lower survival is the curve with `Extent` E_2, highest `Nancy score` and highest `Mayo score`.

The estimated survival function for the time until a patient has an episode of acute severe colitis with the effect of `Involvement beyond endoscopic margin` instead of `Nancy score` is shown in Figure 11.

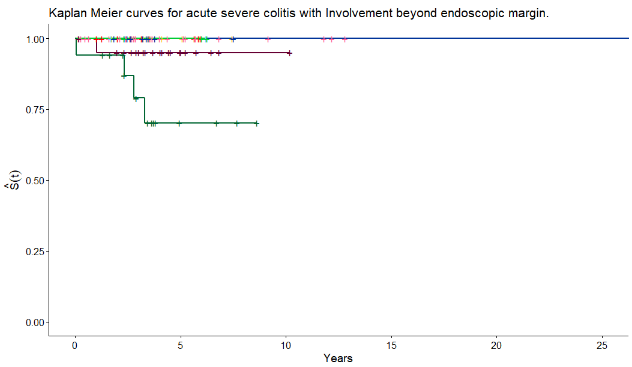


Figure 11: Kaplan-Meier estimate of the survival function for time to acute severe colitis, with the Involvement beyond the margin. • Extent E_1, without Involvement and Low Mayo Score • Extent E_1, without Involvement and High Mayo Score • Extent E_1, with Involvement and Low Mayo Score • Extent E_1, with Involvement and High Mayo Score • Extent E_2, without Involvement and Low Mayo Score • Extent E_2, without Involvement and High Mayo Score • Extent E_2, with Involvement and Low Mayo Score • Extent E_2, with Involvement and High Mayo Score

Four out of five patients belong to the worst strata. This means that the curve with lower survival is the curve that corresponds to the `Extent` E_2, higher `Mayo score` and with `Involvement beyond the endoscopic margin`.

The last Kaplan-Meier curves represent the time until a patient has an episode of acute severe colitis with the effect of `Histological severity` instead of `Nancy score`.

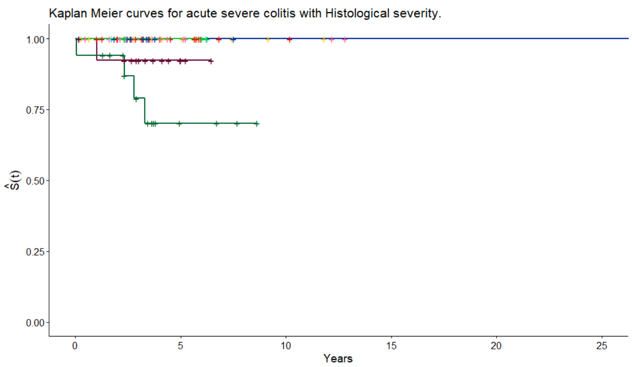


Figure 12: Kaplan-Meier estimate of the survival function for time to acute severe colitis, with the Histological severity. ● Extent E_1, without Histological severity and Low Mayo Score ● Extent E_1, without Histological severity and High Mayo Score ● Extent E_1, with Histological severity and Low Mayo Score ● Extent E_1, with Histological severity and High Mayo Score ● Extent E_2, without Histological severity and Low Mayo Score ● Extent E_2, without Histological severity and High Mayo Score ● Extent E_2, with Histological severity and Low Mayo Score ● Extent E_2, with Histological severity and High Mayo Score

Again, four out of five patients belong to the worst strata. This means that the curve with lower survival corresponds to the worst combination of the categories of the variables.

Colectomy

Considering the need of a colectomy, we obtained 1 observed time and 92 censored times.

In Figure 13, a histogram is shown with the only observed time and with the censored times for the need for a colectomy.

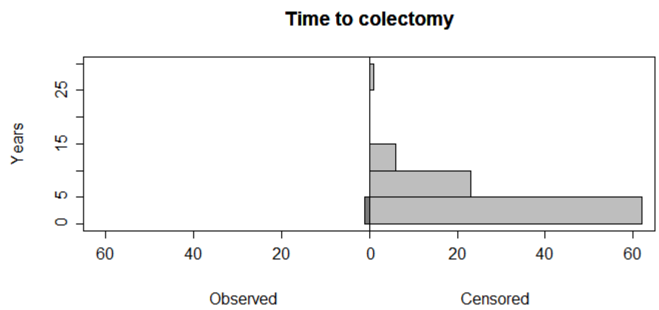


Figure 13: Histogram of observed and censored times to colectomy.

Only one person needed a colectomy and the observed time to this outcome occurred at 2.80 years. The mean time to the censored time was 5.92 years, with a minimum of 0.18 and a maximum of 26.95 years.

The three Kaplan-Meier curves, Figures 14, 15 and 16, have the same conclusion. The only individual who needed to be colectomized corresponds to the most severe combination of variables, whether it is considered Nancy score, `Involvement beyond endoscopic margin` and `Histological severity`.

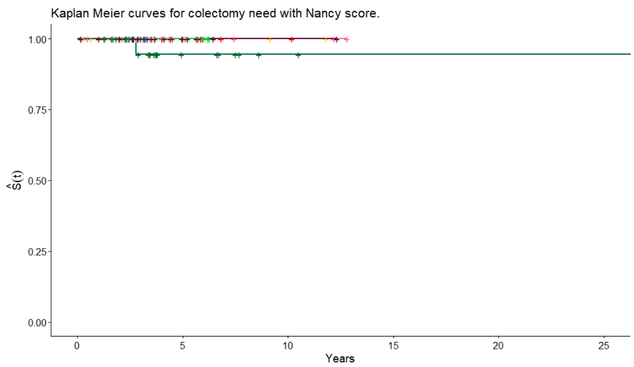


Figure 14: Kaplan-Meier estimate of the survival function for time to colectomy. ● Extent E_1, Nancy < 3 and Low Mayo Score ● Extent E_1, Nancy < 3 and High Mayo Score ● Extent E_1, Nancy ≥ 3 and Low Mayo Score ● Extent E_1, Nancy ≥ 3 and High Mayo Score ● Extent E_2, Nancy < 3 and Low Mayo Score ● Extent E_2, Nancy < 3 and High Mayo Score ● Extent E_2, Nancy ≥ 3 and Low Mayo Score ● Extent E_2, Nancy ≥ 3 and High Mayo Score

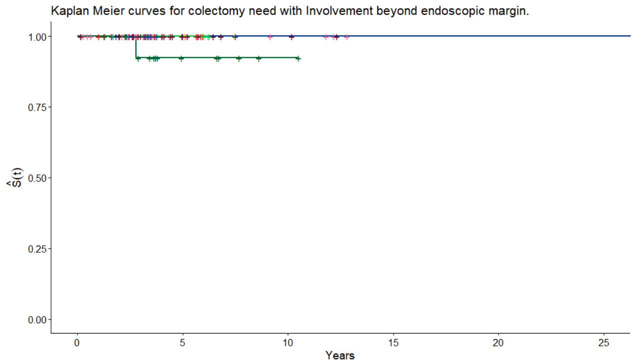


Figure 15: Kaplan-Meier estimate of the survival function for time to colectomy, with Involvement beyond margin. ● Extent E_1, without Involvement and Low Mayo Score ● Extent E_1, without Involvement and High Mayo Score ● Extent E_1, with Involvement and Low Mayo Score ● Extent E_1, with Involvement and High Mayo Score ● Extent E_2, without Involvement and Low Mayo Score ● Extent E_2, without Involvement and High Mayo Score ● Extent E_2, with Involvement and Low Mayo Score ● Extent E_2, with Involvement and High Mayo Score

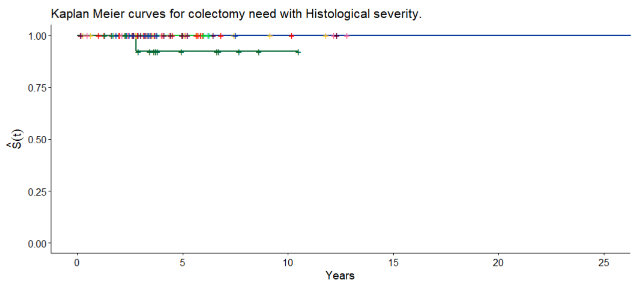


Figure 16: Kaplan-Meier estimate of the survival function for time to colectomy, with Histological severity. ● Extent E_1, without Histological severity and Low Mayo Score ● Extent E_1, without Histological severity and High Mayo Score ● Extent E_1, with Histological severity and Low Mayo Score ● Extent E_1, with Histological severity and High Mayo Score ● Extent E_2, without Histological severity and Low Mayo Score ● Extent E_2, without Histological severity and High Mayo Score ● Extent E_2, with Histological severity and Low Mayo Score ● Extent E_2, with Histological severity and High Mayo Score

Dysplasia

Considering the dysplasia outcome, there were obtained 2 observed times and 91 censored times.

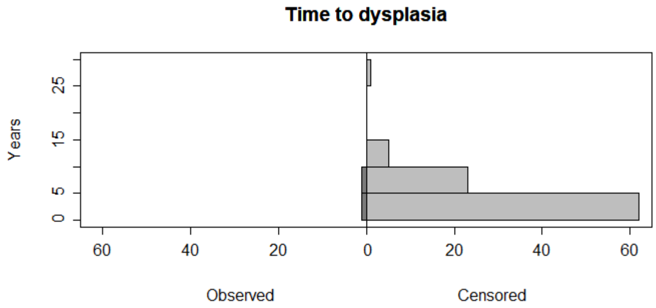


Figure 17: Histogram of observed and censored times to dysplasia.

Figure 17 illustrates the distribution of observed and censored times to dysplasia outcome. For the patients that had a dysplasia, the times were between 0.12 and 6.27 years with a mean time of 3.19 years. For the censored ones, they have quite superior times. The mean of censored times is 4.43 years, with a minimum of 0.18 years and a maximum of 26.95 years.

It will be presented Kaplan-Meier curves to the models initially proposed. The models will not be fitted since there is only two individuals having the outcome.

The estimated survival function for the time until a patient has a dysplasia with the effect of `Extent`, `Nancy score` and `Mayo score` is shown in Figure 18.

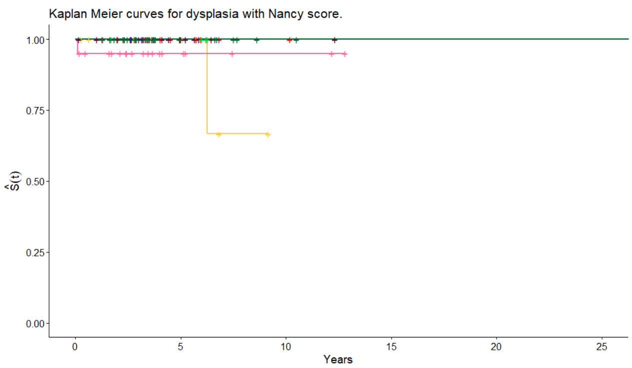


Figure 18: Kaplan-Meier estimate of the survival function for time to dysplasia. • Extent E_1, Nancy < 3 and Low Mayo Score • Extent E_1, Nancy < 3 and High Mayo Score • Extent E_1, Nancy ≥ 3 and Low Mayo Score • Extent E_1, Nancy ≥ 3 and High Mayo Score • Extent E_2, Nancy < 3 and Low Mayo Score • Extent E_2, Nancy < 3 and High Mayo Score • Extent E_2, Nancy ≥ 3 and Low Mayo Score • Extent E_2, Nancy ≥ 3 and High Mayo Score

The estimated survival function for the time until a patient has a dysplasia with the effect of `Extent`, `Involvement beyond endoscopic margin` and `Mayo score` is shown in Figure 19.

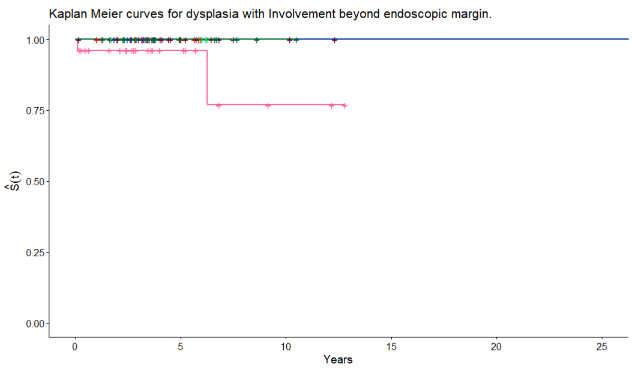


Figure 19: Kaplan-Meier estimate of the survival function for time to dysplasia, with Involvement beyond margin. ● Extent E_1, without Involvement and Low Mayo Score ● Extent E_1, without Involvement and High Mayo Score ● Extent E_1, with Involvement and Low Mayo Score ● Extent E_1, with Involvement and High Mayo Score ● Extent E_2, without Involvement and Low Mayo Score ● Extent E_2, without Involvement and High Mayo Score ● Extent E_2, with Involvement and Low Mayo Score ● Extent E_2, with Involvement and High Mayo Score

The two patients with real lifetimes belong both to the strata with less extent, with involvement beyond margin and with low Mayo score.

Figure 20 represents the Kaplan-Meier curves to the model that includes the `Extent`, `Mayo score` and `Histological severity`.

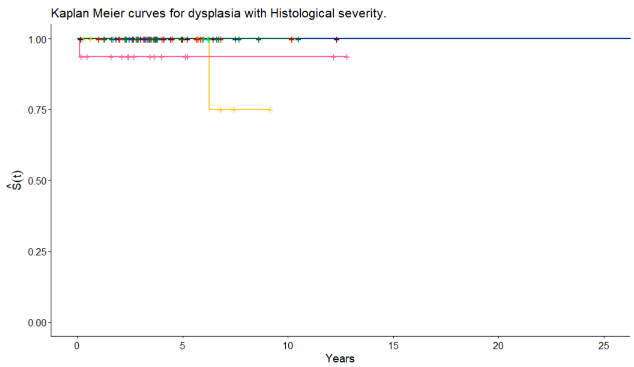


Figure 20: Kaplan-Meier estimate of the survival function for time to dysplasia, with Histological severity. ● Extent E_1, without Histological severity and Low Mayo Score ● Extent E_1, without Histological severity and High Mayo Score ● Extent E_1, with Histological severity and Low Mayo Score ● Extent E_1, with Histological severity and High Mayo Score ● Extent E_2, without Histological severity and Low Mayo Score ● Extent E_2, without Histological severity and High Mayo Score ● Extent E_2, with Histological severity and Low Mayo Score ● Extent E_2, with Histological severity and High Mayo Score

After analyzing the simple outcome models, composite outcome models were created.

Composed outcomes - Therapeutic needs

For the therapeutic needs we obtained 21 observed times and 72 censored times.

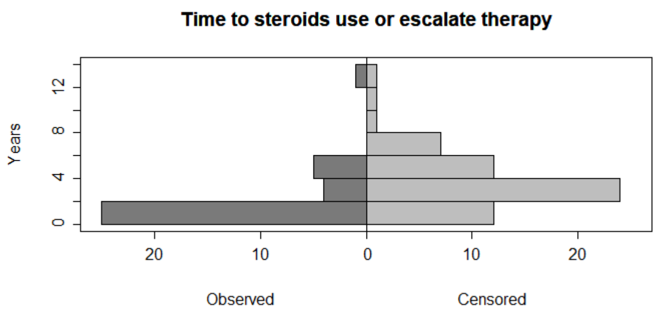


Figure 21: Histogram of observed and censored times to therapeutical needs.

Figure 21 represents the distribution of observed and censored times to the therapeutic needs. It can be seen that the individuals that needed therapeutics have times between 0.04 and 12.18 years, with a mean time of 2.24 years. The mean of censored times is 4.05 years, with a minimum of 0.18 years and a maximum of 12.82 years.

In Table 9 it is represented the six univariates models to predict the hazard for escalate therapy or steroids use.

Table 9: Cox regression univariate analysis for outcome steroids use or escalate therapy.

Variable	β	hazard ratio	p-value
Extent - E$_2$	0.390	1.476	0.378
Age	-0.014	0.986	0.332
Nancy score $\geq$ 3	0.688	1.989	0.274
Mayo score - Moderate and Severe	0.584	1.794	0.231
Involvement beyond margin - Yes	1.020	2.772	0.102
Histological severity	0.957	2.605	0.050

As shown in Table 9, the variable that is significantly associated to the therapeutic needs is `Histological severity`. For an individual with histological severity the hazard of therapeutic needs increases 1.605 compared with a patient without histological severity.

Considering `Extent`, `Nancy score` and `Mayo score` as covariates, the Kaplan-Meier curves for the combination of these categorical variables are presented in Figure 22.

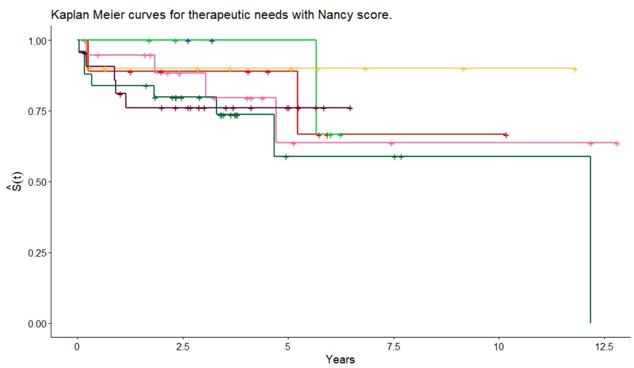


Figure 22: Kaplan-Meier estimate of the survival function for time to therapeutic needs. • Extent E$_1$, Nancy $<$ 3 and Low Mayo Score • Extent E$_1$, Nancy $<$ 3 and High Mayo Score • Extent E$_1$, Nancy $\geq$ 3 and Low Mayo Score • Extent E$_1$, Nancy $\geq$ 3 and High Mayo Score • Extent E$_2$, Nancy $<$ 3 and Low Mayo Score • Extent E$_2$, Nancy $<$ 3 and High Mayo Score • Extent E$_2$, Nancy $\geq$ 3 and Low Mayo Score • Extent E$_2$, Nancy $\geq$ 3 and High Mayo Score

The curve with lower survival is the curve that corresponds to the individuals with the strata: `Extent` E$_2$, `Nancy` $\geq$ 3 and the higher `Mayo score`. This means that the combination of the most severe categories of the variables leads to a lower survival.

There was implemented the model that includes the categorical variables presented above and `Age`. The results of this Cox model are presented in Table 10.

Table 10: Cox regression multiple analysis for outcome steroids use or escalate therapy, with Nancy score as a predictor.

Variable	β	hazard ratio	p-value
Extent - E$_2$	0.081	1.085	0.865
Age	-0.014	0.986	0.320
Nancy score $\geq$ 3	0.673	1.960	0.305
Mayo score - Moderate and Severe	0.477	1.611	0.351

In Table 10, it is possible to notice that there are no significant variables.

Then, the estimated survival function for the time until the use of steroids and escalate therapy with the effect of `Involvement beyond the endoscopic margin` instead of `Nancy` score are shown in Figure 23.

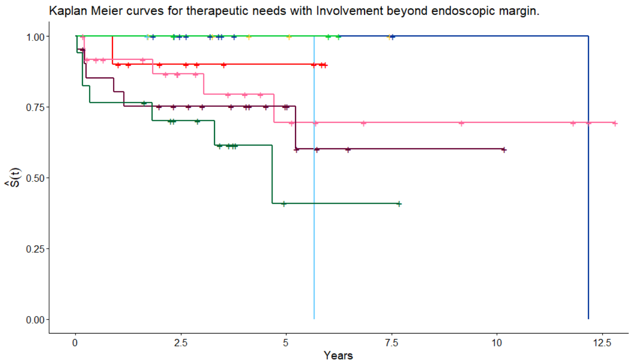


Figure 23: Kaplan-Meier estimate of the survival function for time to therapeutic needs, with Involvement beyond endoscopic margin. ● Extent E_1, without Involvement and Low Mayo Score ● Extent E_1, without Involvement and High Mayo Score ● Extent E_1, with Involvement and Low Mayo Score ● Extent E_1, with Involvement and High Mayo Score ● Extent E_2, without Involvement and Low Mayo Score ● Extent E_2, without Involvement and High Mayo Score ● Extent E_2, with Involvement and Low Mayo Score ● Extent E_2, with Involvement and High Mayo Score

The results of the fitted model that included the variables Extent, Age, Involvement beyond the endoscopic margin and Mayo score are presented in Table 11.

Table 11: Cox regression multiple analysis for outcome steroids use or escalate therapy, with Involvement beyond margin as a predictor.

Variable	β	hazard ratio	p-value
Extent - E_2	0.402	1.495	0.392
Age	-0.015	0.986	0.311
Involvement beyond margin - Yes	1.296	3.653	0.045
Mayo score - Moderate and Severe	0.655	1.924	0.212

Involvement beyond the endoscopic margin is the only variable significantly associated with the need of therapeutics. For a patient with involvement beyond the margin, the hazard of escalate therapy or steroids use is 2.653 higher comparing to a patient without involvement beyond endoscopic margin, by the other variables.

The last model is the one where is included the variable Histological severity instead of Nancy score.

First it was produced the Kaplan-Meier curves for the categorical variables of this model.

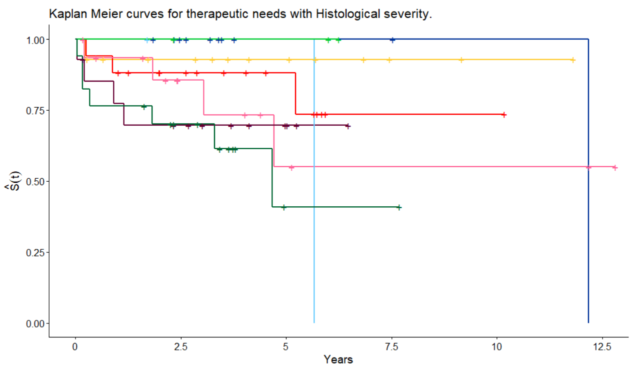


Figure 24: Kaplan-Meier estimate of the survival function for time to therapeutic needs, with Histological severity. • Extent E_1, without Histological severity and Low Mayo Score • Extent E_1, without Histological severity and High Mayo Score • Extent E_1, with Histological severity and Low Mayo Score • Extent E_1, with Histological severity and High Mayo Score • Extent E_2, without Histological severity and Low Mayo Score • Extent E_2, without Histological severity and High Mayo Score • Extent E_2, with Histological severity and Low Mayo Score • Extent E_2, with Histological severity and High Mayo Score

Table 12: Cox regression multiple analysis for outcome steroids use or escalate therapy, with Histological severity, in patients.

Variable	β	hazard ratio	p-value
Extent - E_2	0.104	1.110	0.824
Age	-0.016	0.984	0.250
Histological severity	1.077	2.935	0.031
Mayo score - Moderate and Severe	0.650	1.916	0.215

The variable that is significantly associated with the steroids use and the escalate therapy need is `Histological severity`.

For an individual that has histological severity, the hazard increases 1.935 when compared with an individual that does not have histological severity, by the other variables.

B Residual Analysis

Schoenfeld residuals

The Schoenfeld residuals to the model presented in Table 2 is presented below.

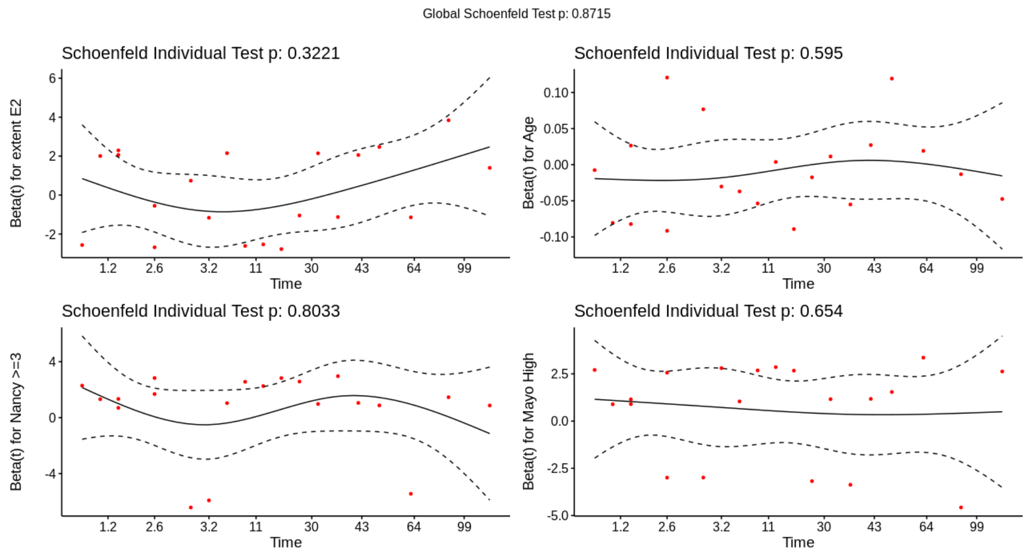


Figure 25: Schoenfeld residuals to multiple model with steroids use outcome with Nancy score included.

From Figure 25 it can be seen that both the variables by themselves with the influence of the rest, as well as all in a global context, are not significant. Considering also the observation of the pattern of the figures, it can be noted that there is no pattern.

Therefore, it can be assumed that the proportional hazard assumption is not violated.

The same analysis was made for the same outcome but considering the variable `Involvement beyond the margin` instead of `Nancy score`. The output is presented in Figure 26.

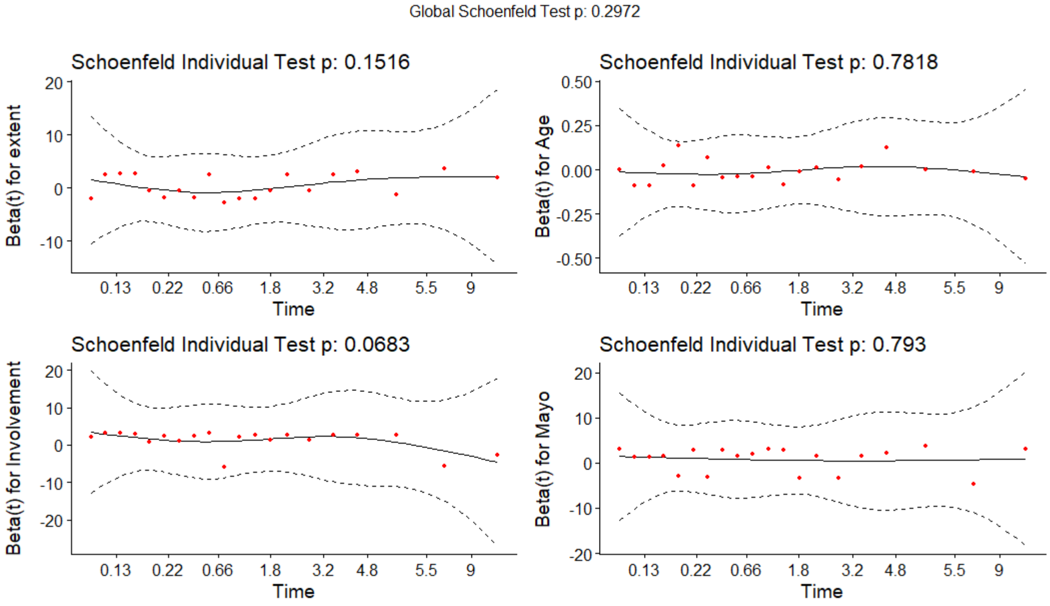


Figure 26: Schoenfeld residuals to multivariate model with steroids use outcome, with Involvement beyond margin included.

The variable `Involvement beyond the endoscopic margin` is significant, but the global p-value is not. This way, the proportional hazards doesn't seem to have been violated.

After this, was made the same to the model that includes the Nancy score and the involvement together. The Schoenfeld residuals are presented next.

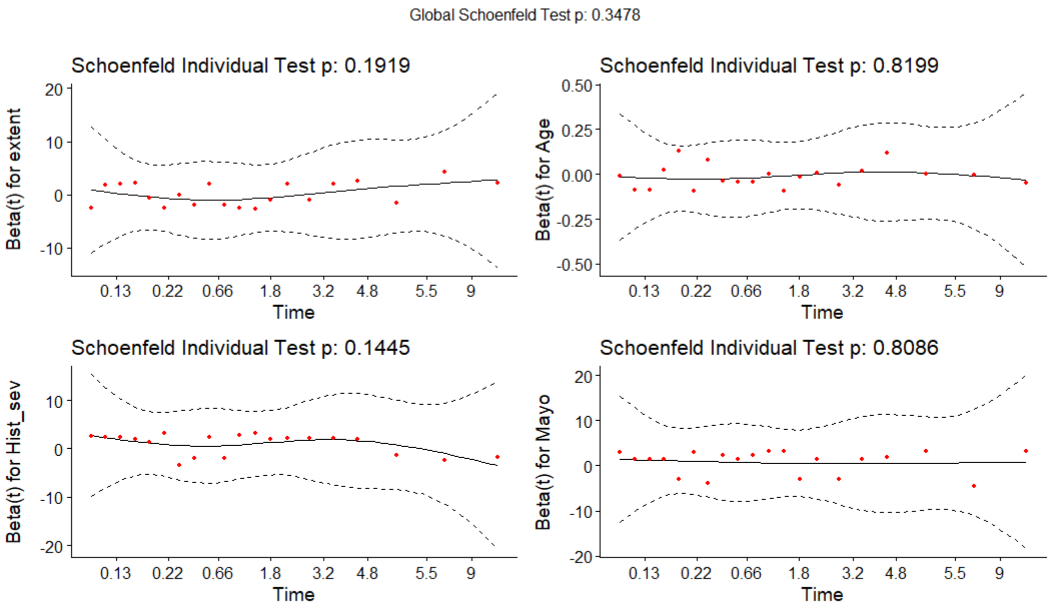


Figure 27: Schoenfeld residuals to multivariate model with steroids use outcome, with Histological severity (Involvement beyond margin and Nancy score combined).

The same analysis was made to the rest of the simple outcomes.

The univariate models to the escalate therapy didn't present a violation to the proportional hazard assumption. There is no pattern of the residuals to each of the covariates and the p-values are all greater than 0.3.

The Schoenfeld residuals to the multiple model, with the Nancy score, are not significant, as we can see in Figure 28.

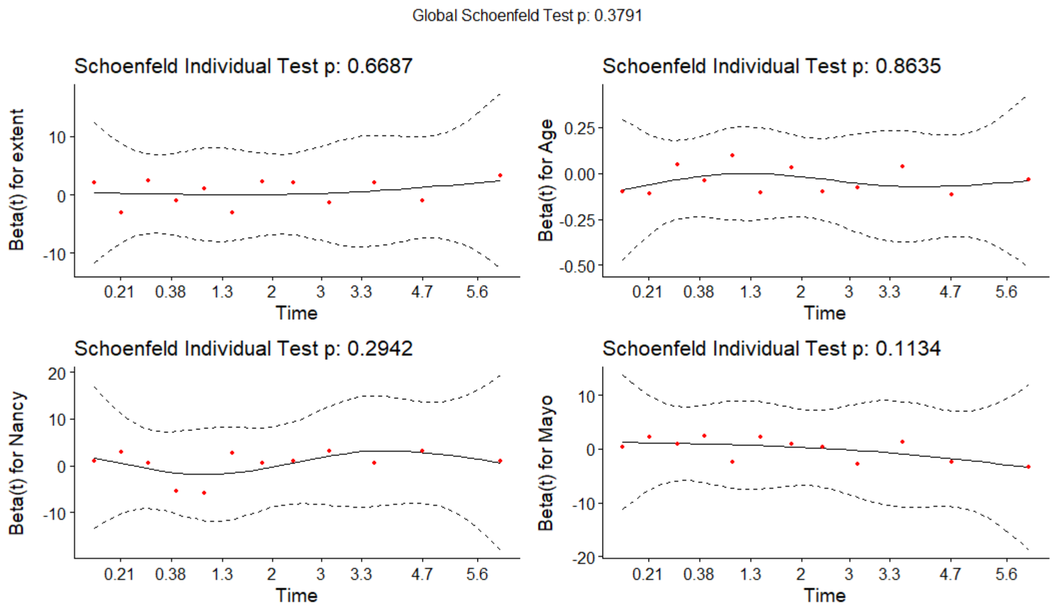


Figure 28: Schoenfeld residuals to multiple model with escalate therapy need.

As it was said before, the p-values are all greater to 0.3 and the global p-value is also not significant. Thus we can conclude that the proportional hazard assumption is not violated.

The same analysis was made for the model that includes the involvement beyond the endoscopic margin instead of Nancy score.

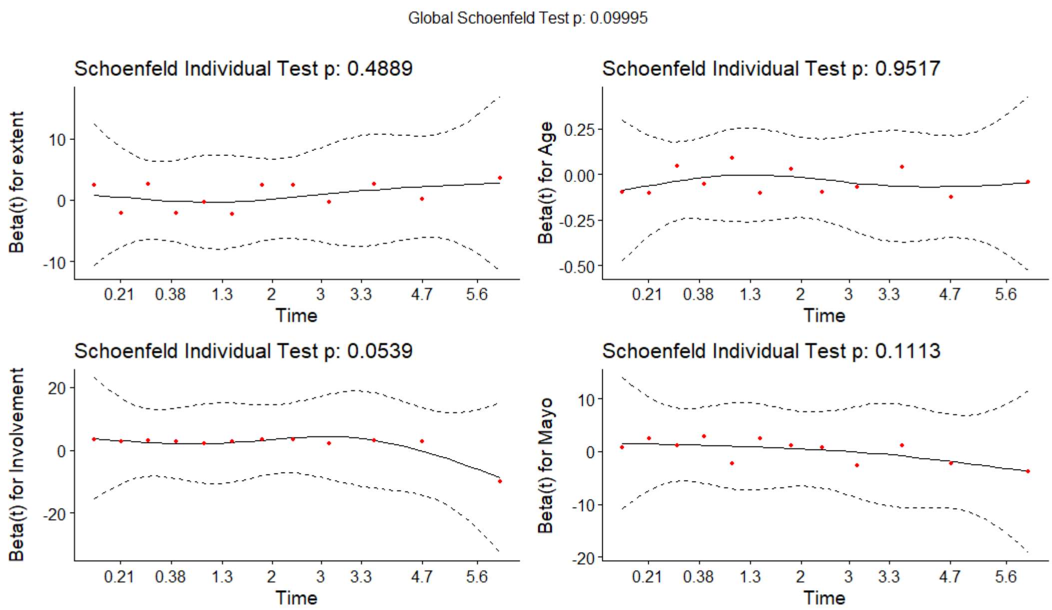


Figure 29: Schoenfeld residuals to multiple model with escalate therapy need, with Involvement beyond margin.

It can be concluded that there is no pattern, and the p-values are not significant. So, the proportional hazards assumption are not violated.

For the model that presents the histological severity, the residuals are:

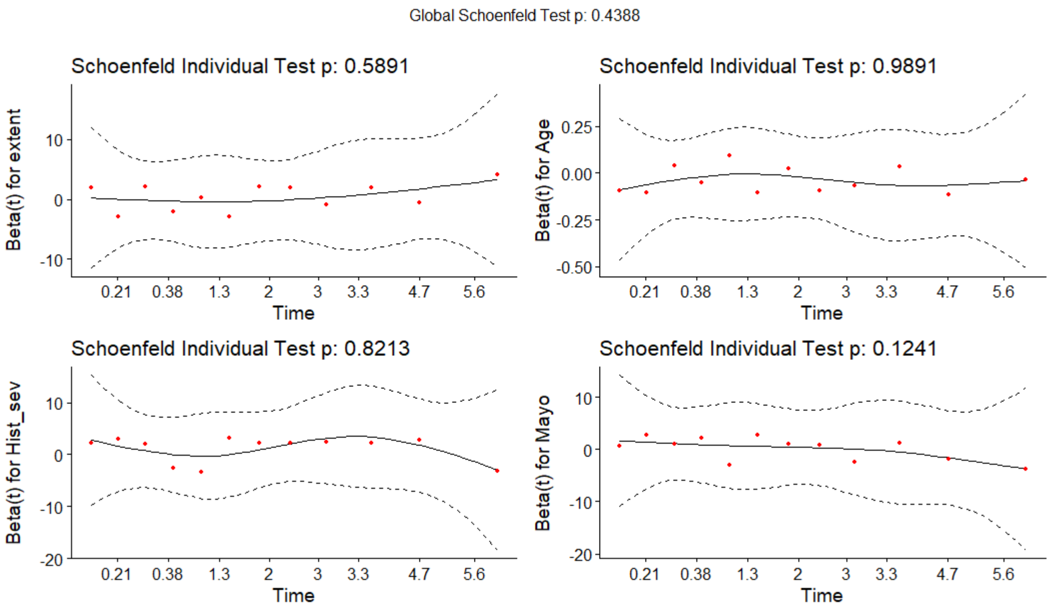


Figure 30: Schoenfeld residuals to multiple model with escalate therapy need, with Histological severity.

After interpreting the Schoenfeld residuals to the simple outcome models, the same analysis was done but this time for the composite outcomes.

Initially the residues were analyzed for the model composed by steroids use and need to escalate the therapy.

Considering this outcome as the response variable and each of the variables as predictors, six univariate models were obtained. For each of these models, the residuals were analyzed and found not to be significant. Therefore, the assumption of proportional hazards for each variable considered separately was not violated. The same can be verified when considering a model that includes the variables with the Nancy score in one model as it is shown in Figure 31.

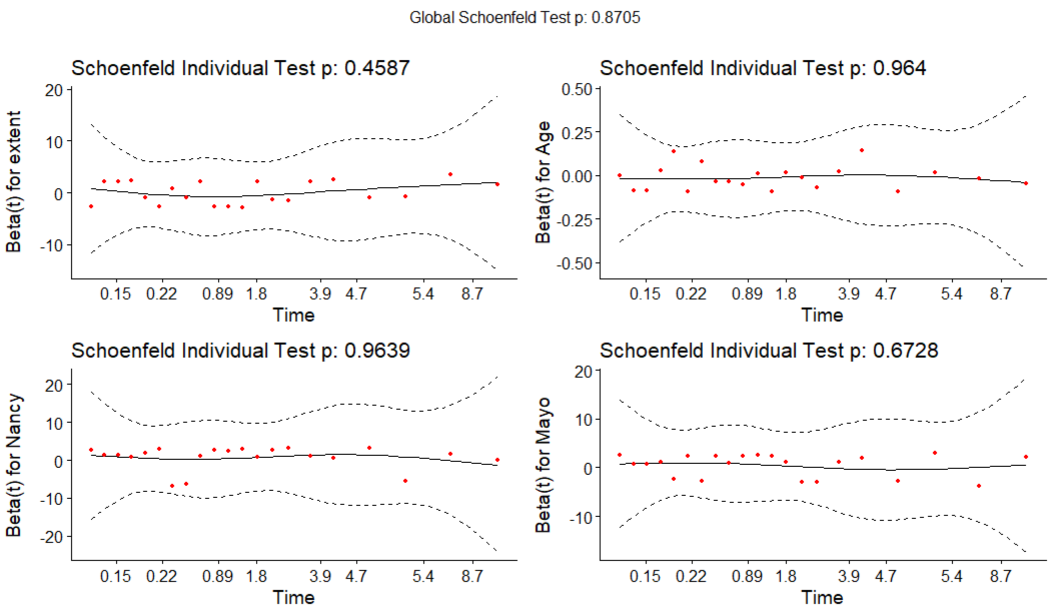


Figure 31: Schoenfeld residuals to multivariate model for steroids use or escalate therapy need.

The p-values for each variable as well as the global p-value for the multivariate model are not significant. This means that we can assume the proportional hazards for this outcome.

Then, considering the model that includes the involvement beyond the endoscopic margin instead of Nancy score, the Schoenfeld residuals are:

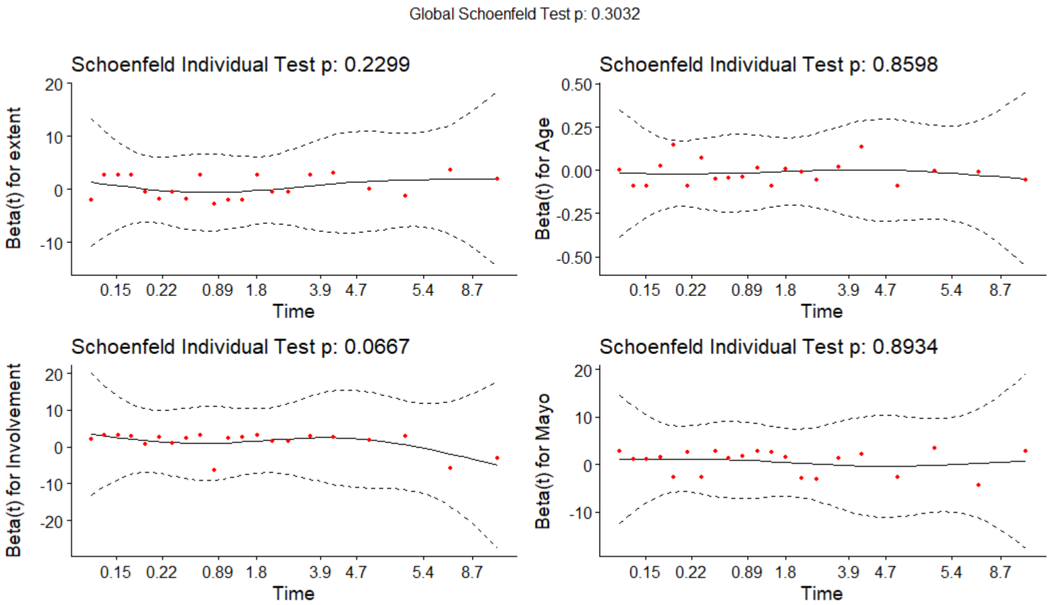


Figure 32: Schoenfeld residuals to multiple model for steroids use or escalate therapy need, with Involvement beyond the endoscopic margin.

Since the global p-value is not significant, the proporional hazard assumption is not violated. The variable involvement need to be inspected.

The Schoenfeld residuals to the model that includes the histological severity are presented below.

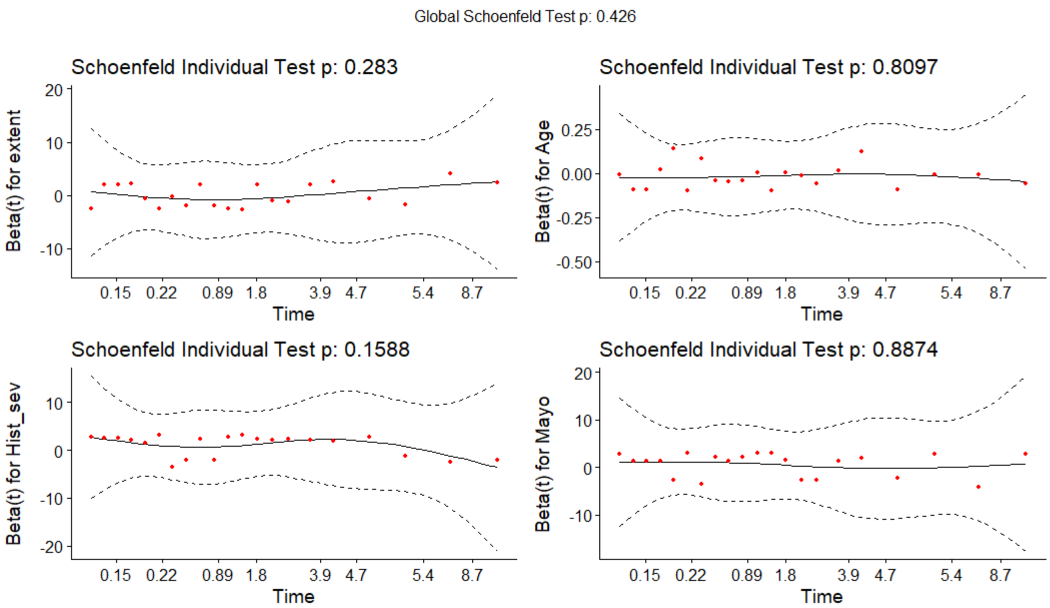
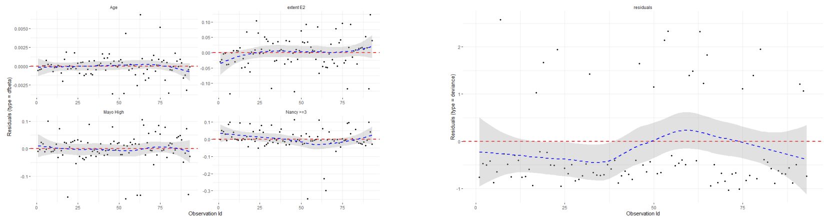


Figure 33: Schoenfeld residuals to multiple model for steroids use or escalate therapy need, with Histological severity.

Since the p-values are not significant, the proportional hazard assumption is not violated.

Dfbeta values and deviance residuals

In the Figure 34 are represented the dfbeta values and the deviance residuals for the simple outcome steroids use, including the Nancy score variable.

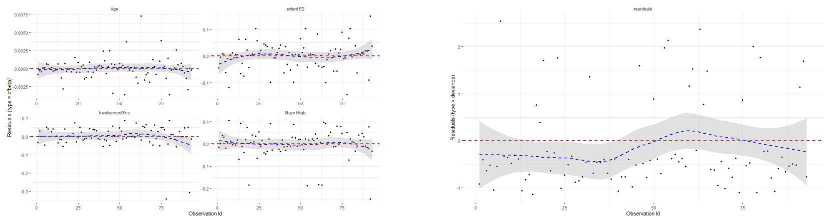


(a) Dfbeta values (b) Deviance residuals

Figure 34: Dfbeta values and deviance residuals to multiple model to steroids use.

Since, from the Figure 34, the pattern looks fairly symmetric around 0, there is no reason to assume that there are influential observations or outliers.

After this, there were reproduced the same analysis but with variable involvement beyond the endoscopic margin instead of Nancy score.

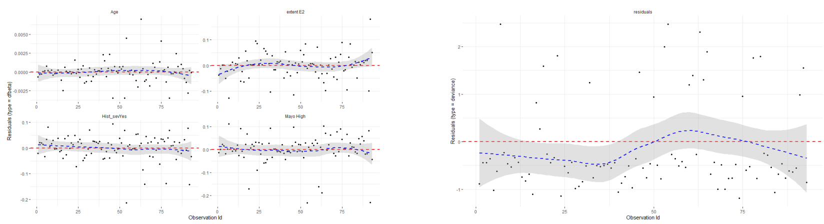


(a) Dfbeta values (b) Deviance residuals

Figure 35: Dfbeta values and deviance residuals to multiple model to steroids use, with Involvement beyond endoscopic margin.

The conclusion for the residuals presented in Figure 35 is basically the same.

Finally, there were produced the dfbeta values and the deviance residuals to the model that includes the extent, age, Mayo score and histological severity, presented in Figure 36.

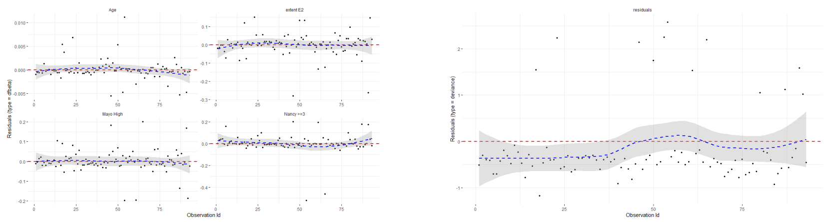


(a) Dfbeta values (b) Deviance residuals

Figure 36: Dfbeta values and deviance residuals to multiple model to steroids use, with Histological severity.

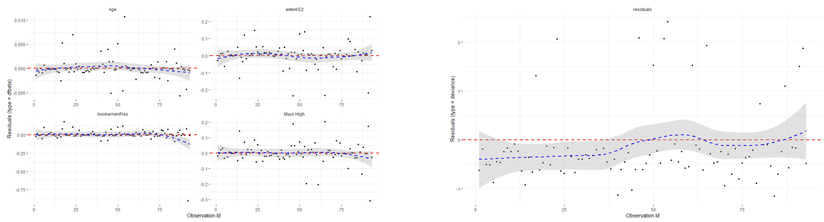
The pattern seems to be symmetric around 0, so there is no reason to assume that there are influential observation or outliers.

This approch was also made to escalate therapy outcome and in the Figures 37 and 38 the residuals are represented.



(a) Dfbeta values (b) Deviance residuals

Figure 37: Dfbeta values and deviance residuals to multivariate model to escalate therapy, with Nancy score.



(a) Dfbeta values (b) Deviance residuals

Figure 38: Dfbeta values and deviance residuals to multiple model to escalate therapy, with Involvement beyond margin.

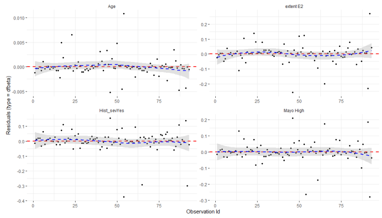
When analysing the figures, it is possible to see a slight trend towards more positive observations, however, this trend does not appear to be significant. Therefore, we can conclude that these observations do not appear to be influential or outliers.

Since there is few people having the outcomes acute severe colitis, colectomy and dysplasia, the residual analysis will not be performed.

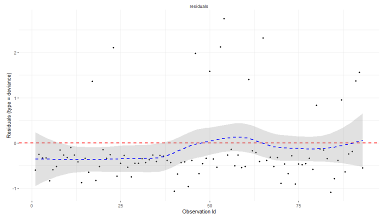
After analyzing the influential observations and outliers of the single outcome models, a similar analysis was performed for the outcomes composed of two outcomes.

The dfbeta values and the deviance residuals were produced to the outcome composed by steroids use and escalate therapy. The plots are presented below.

According to Figures 40, 41 and 42, the pattern of this plots are relatively symmetrical around 0. With this, it can be concluded that there are no influential observations or outliers.

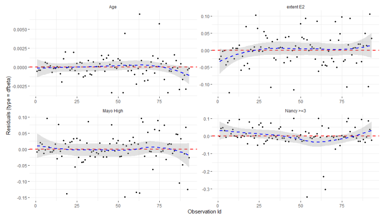


(a) Dfbeta values

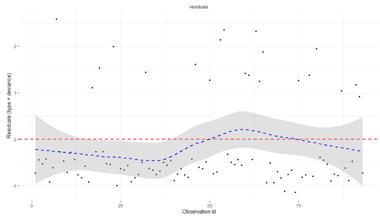


(b) Deviance residuals

Figure 39: Dfbeta values and deviance residuals to multiple model to escalate therapy, with Histological severity.

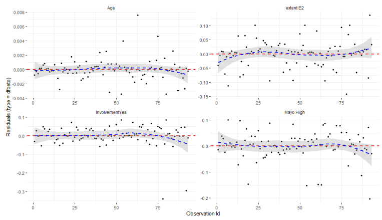


(a) Dfbeta values

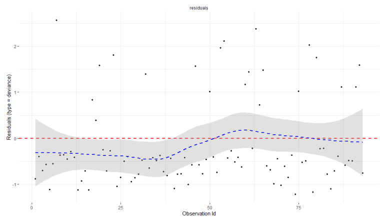


(b) Deviance residuals

Figure 40: Dfbeta values and deviance residuals to multivariate model to steroids use and escalate therapy.

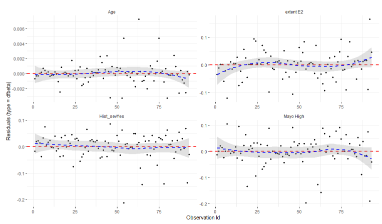


(a) Dfbeta values

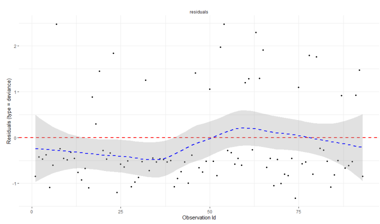


(b) Deviance residuals

Figure 41: Dfbeta values and deviance residuals to multivariate model to steroids use and escalate therapy, with Involvement beyond margin.



(a) Dfbeta values



(b) Deviance residuals

Figure 42: Dfbeta values and deviance residuals to multivariate model to steroids use and escalate therapy, with Histological severity.

Annex

A Data Dictionary Codebook

The Predictive value of Microscopic Inflammation beyond The Endoscopic Margin at diagnosis, in Ulcerative Colitis Outcomes

Codebook

Data Dictionary Codebook

18.10.2018 19:13

#	Variable / Field Name	Field Label *Field Note*	Field Attributes (Field Type, Validation, Choices, Calculations, etc.)
Instrument: **Inclusion Criteria**			
1	record_id	Record ID	text
2	criteria_uc_dx	Section Header: *Inclusion Criteria (all must be met)* Formal diagnosis of Ulcerative Colitis	checkbox 1 criteria_uc_dx___1
3	criteria_proctitis	Proctitis or left sided colitis at diagnosis based on Montreal Classification	checkbox 1 criteria_proctitis___1
4	criteria_histology	Index histology slides from at least 3 different colonic segments are available for review	checkbox 1 criteria_histology___1
5	criteria_info	Detailed information on charts that allows for collection of outcomes	checkbox 1 criteria_info___1
6	criteria_endoscopy	The index endoscopy (endoscopy at diagnosis) is available	checkbox 1 criteria_endoscopy___1
7	criteria_pancolitis	Section Header: *Exlusion Criteria* Endoscopic extensive colitis or pancolitis at diagnosis	checkbox 1 criteria_pancolitis___1
8	criteria_info_followup	Patient does not have a detailed clinical information during the follow-up	checkbox 1 criteria_info_followup___1
9	criteria_biopsies	Biopsies made only after initiating therapy	checkbox 1 criteria_biopsies___1
10	exclusion_note Show the field ONLY if: ([criteria_uc_dx(1)] = '' or [criteria_proctitis(1)] = '' or [criteria_histology(1)] = '' or [criteria_info(1)] = '' or [criteria_endoscopy(1)] = '') or ([criteria_pancolitis(1)] = '1' or [criteria_info_followup(1)] = '1' or [criteria_biopsies(1)] = '1')	The patient does not meet the inclusion criteria.	descriptive
11	inclusion_note Show the field ONLY if: ([criteria_uc_dx(1)] = '1' and [criteria_proctitis(1)] = '1' and [criteria_histology(1)] = '1' and [criteria_info(1)] = '1' and [criteria_endoscopy(1)] = '1') and ([criteria_pancolitis(1)] = '' and [criteria_info_followup(1)] = '' and [criteria_biopsies(1)] = '')	The patient meets the inclusion criteria.	descriptive

12	inclusion_criteria_complete	Section Header: *Form Status* Complete?	dropdown 0 Incomplete 1 Unverified 2 Complete
Instrument: **Patient Information at Diagnosis**			
13	mrn	Medical record number	text, Identifier
14	dob	Date of birth	text (date_dmy, Min: 1900-01-01, Max: 2099-12-31) Field Annotation: @HIDEBUTTON
15	sex	Sex	radio 1 Male 2 Female
16	race	Race	radio 1 White 2 Black/African American 3 Asian 4 American Indian/Alaskan Native 5 Native Hawaiian/Pacific Islander 999 Unknown 888 Other
17	ethnicity	Ethnicity	radio 1 Jewish 2 Hispanic 3 Unknown
18	family_history	Family History of IBD?	radio 1 Yes 2 No 999 Unknown
19	smoking	Smoking habits at diagnosis	radio 1 Smoker 2 Ex-smoker 3 Never-smoker 999 Unknown
20	extra_intest	Extra-intestinal manifestations at diagnosis or during follow-up	yesno 1 Yes 0 No

21	extra_intest_spec Show the field ONLY if: [extra_intest] = '1'	If yes, please choose the appropriate	checkbox 1 extra_intest_spec___1 Peripheral arthritis 2 extra_intest_spec___2 Ankylosing spondylitis 3 extra_intest_spec___3 Erythema nodosum 4 extra_intest_spec___4 Pyoderma gangrenosum 5 extra_intest_spec___5 Uveitis 6 extra_intest_spec___6 Episcleritis 7 extra_intest_spec___7 Primary sclerosing cholangitis 888 extra_intest_spec___888 Other
22	patient_information_at_diagnosis_complete	Section Header: *Form Status* Complete?	dropdown 0 Incomplete 1 Unverified 2 Complete
Instrument: **Clinical and Endoscopic Information at Diagnosis**			
23	diag_date	Date of diagnosis	text (date_dmy, Min: 1900-01-01, Max: 2099-12-31)
24	diag_age	Age at diagnosis *(automatically calculated)*	calc Calculation: rounddown(datediff([dob], [diag_date],"y","dmy",true))
25	diag_age_categ	Please select diagnosis age category	radio 1 Below 16 years 2 Between 17-40 3 Above 40
26	clinical_severity	Clinical severity at diagnosis *Mild UC - Passage of ≤ 4 stools/day (with or without blood), absence of any systemic illness, and normal inflammatory markers (ESR)Moderate UC - Passage of > 4 stools per day but with minimal signs of systemic toxicity Severe UC Passage of ≥ 6 bloody stools daily, pulse rate of at least 90 bpm, temperature ≥ 37.5°C, haemoglobin < than 10.5 g/100 ml, and ESR ≥ 30 mm/h*	radio 1 Mild UC 2 Moderate UC 3 Severe UC
27	extent	Extent at diagnosis *E1: endoscopic inflammation limited to the rectum E2: endoscopic inflammation up to the splenic flexure*	radio 1 E1 Ulcerative proctitis 2 E2 Left sided UC (distal UC)
28	endo_inflam	Please specify the segments with endoscopic (macroscopic) inflammation	checkbox 1 endo_inflam___1 Rectum 2 endo_inflam___2 Sigmoid 3 endo_inflam___3 Descending colon 4 endo_inflam___4 Transverse colon 5 endo_inflam___5 Ascending colon 6 endo_inflam___6 Cecum 7 endo_inflam___7 Terminal ileum

29	severity	Endoscopic severity at diagnosis *Mild disease - erythema, decreased vascular pattern, mild friability Moderate disease - marked erythema, absent vascular pattern, friability, erosion Severe disease - spontaneous bleeding and ulceration*	radio 1 Normal/quiescent 2 Mild disease 3 Moderate disease 4 Severe disease
30	ulcerations	Presence of deep and extensive ulcerations covering more than 10% of the mucosal area of at least one segment of the colon	yesno 1 Yes 0 No
31	tx_diag	Treatment at diagnosis Please choose the treatment that was started within the first three months after the diagnosis	checkbox 1 tx_diag___1 5-ASA (oral) 2 tx_diag___2 5-ASA (rectal) 3 tx_diag___3 Azathioprine 4 tx_diag___4 6-MP 5 tx_diag___5 Cyclosporine 6 tx_diag___6 Methotrexate 7 tx_diag___7 Infliximab 8 tx_diag___8 Adalimumab 9 tx_diag___9 Certolizumab 10 tx_diag___10 Golimumab 11 tx_diag___11 Vedolizumab 12 tx_diag___12 Oral Steroids 13 tx_diag___13 IV Steroids 14 tx_diag___14 Rectal Steroids 888 tx_diag___888 Other
32	clinical_and_endoscopic_information_at_diagnosis_complete	Section Header: *Form Status* Complete?	dropdown 0 Incomplete 1 Unverified 2 Complete
Instrument: **Follow-up of UC**			
33	followup_date	Date of the last follow-up *(please insert the last date the patient was seen at your institution)*	text (date_dmy, Min: 1900-01-01, Max: 2099-12-31)
34	followup_days	Days of follow-up *(automatically calculated)*	calc Calculation: rounddown(datediff([diag_date], [followup_date], "d","dmy",true))
35	hospit_followup	Need for hospitalisation at diagnosis or during the follow-up	yesno 1 Yes 0 No Custom alignment: RH

36	hospit_number Show the field ONLY if: [hospit_followup] = '1'	Number of hospitalizations	dropdown 1 1 2 2 3 3 4 4 5 5 6 6 7 7 8 8
37	hospit_1 Show the field ONLY if: [hospit_number] = '8' or [hospit_number] = '7' or [hospit_number] = '6' or [hospit_number] = '5' or [hospit_number] = '4' or [hospit_number] = '3' or [hospit_number] = '2' or [hospit_number] = '1'	Date of 1st hospitalization	text (date_dmy)
38	hospit_2 Show the field ONLY if: [hospit_number] = '8' or [hospit_number] = '7' or [hospit_number] = '6' or [hospit_number] = '5' or [hospit_number] = '4' or [hospit_number] = '3' or [hospit_number] = '2'	Date of 2nd hospitalization	text (date_dmy)
39	hospit_3 Show the field ONLY if: [hospit_number] = '8' or [hospit_number] = '7' or [hospit_number] = '6' or [hospit_number] = '5' or [hospit_number] = '4' or [hospit_number] = '3'	Date of 3rd hospitalization	text (date_dmy)
40	hospit_4 Show the field ONLY if: [hospit_number] = '8' or [hospit_number] = '7' or [hospit_number] = '6' or [hospit_number] = '5' or [hospit_number] = '4'	Date of 4th hospitalization	text (date_dmy)
41	hospit_5 Show the field ONLY if: [hospit_number] = '8' or [hospit_number] = '7' or [hospit_number] = '6' or [hospit_number] = '5'	Date of 5th hospitalization	text (date_dmy)
42	hospit_6 Show the field ONLY if: [hospit_number] = '8' or [hospit_number] = '7' or [hospit_number] = '6'	Date of 6th hospitalization	text (date_dmy)
43	hospit_7 Show the field ONLY if: [hospit_number] = '8' or [hospit_number] = '7'	Date of 7th hospitalization	text (date_dmy)

44	hospit_8 Show the field ONLY if: [hospit_number] = '8'	Date of 8th hospitalization	text (date_dmy)
45	colectomy	Colectomy at diagnosis or during follow-up	yesno 1 Yes 0 No
46	colectomy_date Show the field ONLY if: [colectomy] = '1'	Date	text (date_dmy, Min: 1900-01-01, Max: 2099-12-31)
47	asuc	Acute severe ulcerative colitis at diagnosis or during follow-up	yesno 1 Yes 0 No
48	asuc_date Show the field ONLY if: [asuc] = '1'	Date	text (date_dmy, Min: 1900-01-01, Max: 2099-12-31)
49	proximal_doc	Proximal extension of the disease documented during follow-up *(defined as a greater extent of macroscopic inflammation than the one observed in the index colonoscopy)*	yesno 1 Yes 0 No Custom alignment: RH
50	proximal_doc_spec Show the field ONLY if: [proximal_doc] = '1'	Please specify (according to Montreal Classification)	radio 1 E1 to E2 2 E1 to E3 3 E2 to E3
51	steroids	Section Header: Steroids used during follow-up *(do not include steroids course that was needed at the moment of diagnosis to induce remission)*	yesno 1 Yes 0 No Custom alignment: RH
52	steroids_number Show the field ONLY if: [steroids] = '1'	Number of courses of steroids used during follow-up	text (integer) Custom alignment: RH
53	steroids_date Show the field ONLY if: [steroids] = '1'	Date of first flare requiring steroids	text (date_dmy, Min: 1900-01-01, Max: 2099-12-31) Custom alignment: RH
54	escalate_thr	Need to escalate therapy during follow-up (after 3 months of follow-up) *(defined as a need to increase therapy or switch drugs within the same class of drug or to introduce a new class of drugs)*	yesno 1 Yes 0 No Custom alignment: RH
55	escalate_thr_5asa Show the field ONLY if: [escalate_thr] = '1'	Increase dose of 5-ASA agents	yesno 1 Yes 0 No
56	escalate_thr_5asa_date Show the field ONLY if: [escalate_thr_5asa] = '1'	Date	text (date_dmy, Min: 1900-01-01, Max: 2099-12-31)
57	escalate_thr_azat Show the field ONLY if: [escalate_thr] = '1'	Azathioprine/ 6-MP	yesno 1 Yes 0 No

58	escalate_thr_azat_date Show the field ONLY if: [escalate_thr_azat] = '1'	Date	text (date_dmy, Min: 1900-01-01, Max: 2099-12-31)
59	escalate_thr_cyclo Show the field ONLY if: [escalate_thr] = '1'	Cyclosporine	yesno 1 Yes 0 No
60	escalate_thr_cyclo_date Show the field ONLY if: [escalate_thr_cyclo] = '1'	Date	text (date_dmy, Min: 1900-01-01, Max: 2099-12-31)
61	escalate_thr_metho Show the field ONLY if: [escalate_thr] = '1'	Methotrexate	yesno 1 Yes 0 No
62	escalate_thr_metho_date Show the field ONLY if: [escalate_thr_metho] = '1'	Date	text (date_dmy, Min: 1900-01-01, Max: 2099-12-31)
63	escalate_thr_bio Show the field ONLY if: [escalate_thr] = '1'	Biologics (anti-TNF or vedolizumab)	yesno 1 Yes 0 No
64	escalate_thr_bio_date Show the field ONLY if: [escalate_thr_bio] = '1'	Date	text (date_dmy, Min: 1900-01-01, Max: 2099-12-31)
65	resp_med	Section Header: During follow-up was the patient considered to be at any point:	descriptive
66	resp_med_steroid_dep	a) steroid-dependent	yesno 1 Yes 0 No
67	resp_med_steroid_refr	b) steroid-refractory	yesno 1 Yes 0 No
68	resp_med_immuno	c) non-respondent to immunomodulator *(do not include as non-response patients that did not tolerate the drug because of side-effects)*	yesno 1 Yes 0 No
69	resp_med_bio_nonresp	d) biologic non-respondent	yesno 1 Yes 0 No
70	resp_med_bio_ref	e) biologic refractory	yesno 1 Yes 0 No
71	followup_dysplasia	Section Header: Colonic mucosal dysplasia evidenced during the follow-up	yesno 1 Yes 0 No

72	followup_dysplasia_spec Show the field ONLY if: [followup_dysplasia] = '1'	Please specify	radio 1 Low grade dysplasia 2 High grade dysplasia 3 Colorectal cancer
73	followup_dysplasia_date Show the field ONLY if: [followup_dysplasia] = '1'	Date	text (date_dmy, Min: 1900-01-01, Max: 2099-12-31)
74	followup_death	Patient died during the disease course	yesno 1 Yes 0 No
75	followup_death_cause Show the field ONLY if: [followup_death] = '1'	Was the cause of dead a complication of the disease?	yesno 1 Yes 0 No
76	followup_death_date Show the field ONLY if: [followup_death] = '1'	Date	text (date_dmy, Min: 1900-01-01, Max: 2099-12-31)
77	followup_of_uc_complete	Section Header: *Form Status* Complete?	dropdown 0 Incomplete 1 Unverified 2 Complete
Instrument: **Histologic Features**			
78	histo_features	Histologic Features We will review every histology slide and characterize the microscopic features according to the Nancy score. Please indicate if biopsies are available at the moment of diagnosis for review and please select below to let us know that the slides have been sent to us.	descriptive
79	slides_sent	Slides have been sent for histological review	yesno 1 Yes 0 No
80	slides_location	We need to trace the location of the biopsies to the patient so we can review the extent of the microscopic inflammation. Please select the most convenient way for you:	radio 1 Upload the pathology record into REDCap. Delete the name of the patient from the record before uploading it 2 Upload the record directly and we will delete the name of the patient from the record once it is uploaded 3 Send the pathology record with the slides by mail and we will anonymize it and upload the record Custom alignment: LV

81	pathology_upload Show the field ONLY if: [slides_location] = '1' or [slides_location] = '2'	Pathology record upload	file
82	histologic_features_complete	Section Header: *Form Status* Complete?	dropdown 0 Incomplete 1 Unverified 2 Complete
Instrument: **Pathology Assessment**			
83	path_id	Pathology ID	text
84	slides_number	Number of slides received	dropdown 1 1 2 2 3 3 4 4 5 5 6 6 7 7 8 8
85	pathology_notes	Notes	notes
86	slide_01 Show the field ONLY if: [slides_number] = '1' or [slides_number] = '2' or [slides_number] = '3' or [slides_number] = '4' or [slides_number] = '5' or [slides_number] = '6' or [slides_number] = '7' or [slides_number] = '8'	Slide 1	descriptive
87	code_01 Show the field ONLY if: [slides_number] = '1' or [slides_number] = '2' or [slides_number] = '3' or [slides_number] = '4' or [slides_number] = '5' or [slides_number] = '6' or [slides_number] = '7' or [slides_number] = '8'	Biopsy code Slide 1	text
88	location_01 Show the field ONLY if: [slides_number] = '1' or [slides_number] = '2' or [slides_number] = '3' or [slides_number] = '4' or [slides_number] = '5' or [slides_number] = '6' or [slides_number] = '7' or [slides_number] = '8'	Location Slide 1	dropdown 1 Rectum 2 Sigmoid 3 Descending colon 4 Transverse colon 5 Ascending colon 6 Cecum 7 Terminal ileum 8 Right colon 9 Left colon

89	nancy_score_01 Show the field ONLY if: [slides_number] = '1' or [slides_number] = '2' or [slides_number] = '3' or [slides_number] = '4' or [slides_number] = '5' or [slides_number] = '6' or [slides_number] = '7' or [slides_number] = '8'	Nancy score Slide 1 *(defined as the quantity of lymphocytes and plasmacytes in the biopsy specimen)*	descriptive
90	inflam_infiltrate_01 Show the field ONLY if: [slides_number] = '1' or [slides_number] = '2' or [slides_number] = '3' or [slides_number] = '4' or [slides_number] = '5' or [slides_number] = '6' or [slides_number] = '7' or [slides_number] = '8'	Chronic inflammatory infiltrate Slide 1 *(defined as the quantity of lymphocytes and plasmacytes in the biopsy specimen)*	radio 0 No increase 1 Mild but unequivocal increase 2 Moderate increase 3 Marked increase
91	neutrophils_epith_01 Show the field ONLY if: [slides_number] = '1' or [slides_number] = '2' or [slides_number] = '3' or [slides_number] = '4' or [slides_number] = '5' or [slides_number] = '6' or [slides_number] = '7' or [slides_number] = '8'	Neutrophils in the epithelium Slide 1	radio 0 None 1 < 50% crypt involved 2 >50% crypt involved
92	ulceration_01 Show the field ONLY if: [slides_number] = '1' or [slides_number] = '2' or [slides_number] = '3' or [slides_number] = '4' or [slides_number] = '5' or [slides_number] = '6' or [slides_number] = '7' or [slides_number] = '8'	Ulceration Slide 1 *(defined as visible epithelial injury and regeneration and/or fibrin and neutrophils and/or tissue granulation)*	radio 0 Absent 1 Present
93	acute_inflam_01 Show the field ONLY if: [slides_number] = '1' or [slides_number] = '2' or [slides_number] = '3' or [slides_number] = '4' or [slides_number] = '5' or [slides_number] = '6' or [slides_number] = '7' or [slides_number] = '8'	Acute inflammatory cell infiltrate Slide 1	radio 0 None 1 Mild 2 Moderate 3 Severe
94	mucin_01 Show the field ONLY if: [slides_number] = '1' or [slides_number] = '2' or [slides_number] = '3' or [slides_number] = '4' or [slides_number] = '5' or [slides_number] = '6' or [slides_number] = '7' or [slides_number] = '8'	Mucin depletion Slide 1	radio 0 None 1 Mild 2 Moderate 3 Severe

95	neutrophils_lamina_01 Show the field ONLY if: [slides_number] = '1' or [slides_number] = '2' or [slides_number] = '3' or [slides_number] = '4' or [slides_number] = '5' or [slides_number] = '6' or [slides_number] = '7' or [slides_number] = '8'	Neutrophils in lamina propria Slide 1	radio 0 None 1 Mild 2 Moderate 3 Severe
96	basal_plasmacyt_01 Show the field ONLY if: [slides_number] = '1' or [slides_number] = '2' or [slides_number] = '3' or [slides_number] = '4' or [slides_number] = '5' or [slides_number] = '6' or [slides_number] = '7' or [slides_number] = '8'	Basal plasmacytosis Slide 1	radio 0 None 1 Mild 2 Moderate 3 Severe
97	serrated_abnor_01 Show the field ONLY if: [slides_number] = '1' or [slides_number] = '2' or [slides_number] = '3' or [slides_number] = '4' or [slides_number] = '5' or [slides_number] = '6' or [slides_number] = '7' or [slides_number] = '8'	Serrated architectural abnormalities Slide 1 *(defined as the presence of dilated crypts showing a scalloped lumen)*	radio 0 None 1 < 5% crypt involved 2 < 50% crypt involved 3 >50% crypt involved
98	notes_01 Show the field ONLY if: [slides_number] = '1' or [slides_number] = '2' or [slides_number] = '3' or [slides_number] = '4' or [slides_number] = '5' or [slides_number] = '6' or [slides_number] = '7' or [slides_number] = '8'	Notes Slide 1	notes
99	slide_02 Show the field ONLY if: [slides_number] = '2' or [slides_number] = '3' or [slides_number] = '4' or [slides_number] = '5' or [slides_number] = '6' or [slides_number] = '7' or [slides_number] = '8'	Slide 2	descriptive
100	code_02 Show the field ONLY if: [slides_number] = '2' or [slides_number] = '3' or [slides_number] = '4' or [slides_number] = '5' or [slides_number] = '6' or [slides_number] = '7' or [slides_number] = '8'	Biopsy code Slide 2	text

101	location_02 Show the field ONLY if: [slides_number] = '2' or [slides_number] = '3' or [slides_number] = '4' or [slides_number] = '5' or [slides_number] = '6' or [slides_number] = '7' or [slides_number] = '8'	Location Slide 2	dropdown 1 Rectum 2 Sigmoid 3 Descending colon 4 Transverse colon 5 Ascending colon 6 Cecum 7 Terminal ileum 8 Right colon 9 Left colon
102	nancy_score_02 Show the field ONLY if: [slides_number] = '2' or [slides_number] = '3' or [slides_number] = '4' or [slides_number] = '5' or [slides_number] = '6' or [slides_number] = '7' or [slides_number] = '8'	Nancy score Slide 2 *(defined as the quantity of lymphocytes and plasmacytes in the biopsy specimen)*	descriptive
103	inflam_infiltrate_02 Show the field ONLY if: [slides_number] = '2' or [slides_number] = '3' or [slides_number] = '4' or [slides_number] = '5' or [slides_number] = '6' or [slides_number] = '7' or [slides_number] = '8'	Chronic inflammatory infiltrate Slide 2 *(defined as the quantity of lymphocytes and plasmacytes in the biopsy specimen)*	radio 0 No increase 1 Mild but unequivocal increase 2 Moderate increase 3 Marked increase
104	neutrophils_epith_02 Show the field ONLY if: [slides_number] = '2' or [slides_number] = '3' or [slides_number] = '4' or [slides_number] = '5' or [slides_number] = '6' or [slides_number] = '7' or [slides_number] = '8'	Neutrophils in the epithelium Slide 2	radio 0 None 1 < 50% crypt involved 2 >50% crypt involved
105	ulceration_02 Show the field ONLY if: [slides_number] = '2' or [slides_number] = '3' or [slides_number] = '4' or [slides_number] = '5' or [slides_number] = '6' or [slides_number] = '7' or [slides_number] = '8'	Ulceration Slide 2 *(defined as visible epithelial injury and regeneration and/or fibrin and neutrophils and/or tissue granulation)*	radio 0 Absent 1 Present
106	acute_inflam_02 Show the field ONLY if: [slides_number] = '2' or [slides_number] = '3' or [slides_number] = '4' or [slides_number] = '5' or [slides_number] = '6' or [slides_number] = '7' or [slides_number] = '8'	Acute inflammatory cell infiltrate Slide 2	radio 0 None 1 Mild 2 Moderate 3 Severe

107	mucin_02 Show the field ONLY if: [slides_number] = '2' or [slides_number] = '3' or [slides_number] = '4' or [slides_number] = '5' or [slides_number] = '6' or [slides_number] = '7' or [slides_number] = '8'	Mucin depletion Slide 2	radio 0 None 1 Mild 2 Moderate 3 Severe
108	neutrophils_lamina_02 Show the field ONLY if: [slides_number] = '2' or [slides_number] = '3' or [slides_number] = '4' or [slides_number] = '5' or [slides_number] = '6' or [slides_number] = '7' or [slides_number] = '8'	Neutrophils in lamina propria Slide 2	radio 0 None 1 Mild 2 Moderate 3 Severe
109	basal_plasmacyt_02 Show the field ONLY if: [slides_number] = '2' or [slides_number] = '3' or [slides_number] = '4' or [slides_number] = '5' or [slides_number] = '6' or [slides_number] = '7' or [slides_number] = '8'	Basal plasmacytosis Slide 2	radio 0 None 1 Mild 2 Moderate 3 Severe
110	serrated_abnor_02 Show the field ONLY if: [slides_number] = '2' or [slides_number] = '3' or [slides_number] = '4' or [slides_number] = '5' or [slides_number] = '6' or [slides_number] = '7' or [slides_number] = '8'	Serrated architectural abnormalities Slide 2 *(defined as the presence of dilated crypts showing a scalloped lumen)*	radio 0 None 1 < 5% crypt involved 2 < 50% crypt involved 3 >50% crypt involved
111	notes_02 Show the field ONLY if: [slides_number] = '2' or [slides_number] = '3' or [slides_number] = '4' or [slides_number] = '5' or [slides_number] = '6' or [slides_number] = '7' or [slides_number] = '8'	Notes Slide 2	notes
112	slide_03 Show the field ONLY if: [slides_number] = '3' or [slides_number] = '4' or [slides_number] = '5' or [slides_number] = '6' or [slides_number] = '7' or [slides_number] = '8'	Slide 3	descriptive
113	code_03 Show the field ONLY if: [slides_number] = '3' or [slides_number] = '4' or [slides_number] = '5' or [slides_number] = '6' or [slides_number] = '7' or [slides_number] = '8'	Biopsy code Slide 3	text

114	location_03 Show the field ONLY if: [slides_number] = '3' or [slides_number] = '4' or [slides_number] = '5' or [slides_number] = '6' or [slides_number] = '7' or [slides_number] = '8'	Location Slide 3	dropdown 1 Rectum 2 Sigmoid 3 Descending colon 4 Transverse colon 5 Ascending colon 6 Cecum 7 Terminal ileum 8 Right colon 9 Left colon
115	nancy_score_03 Show the field ONLY if: [slides_number] = '3' or [slides_number] = '4' or [slides_number] = '5' or [slides_number] = '6' or [slides_number] = '7' or [slides_number] = '8'	Nancy score Slide 3 *(defined as the quantity of lymphocytes and plasmacytes in the biopsy specimen)*	descriptive
116	inflam_infiltrate_03 Show the field ONLY if: [slides_number] = '3' or [slides_number] = '4' or [slides_number] = '5' or [slides_number] = '6' or [slides_number] = '7' or [slides_number] = '8'	Chronic inflammatory infiltrate Slide 3 *(defined as the quantity of lymphocytes and plasmacytes in the biopsy specimen)*	radio 0 No increase 1 Mild but unequivocal increase 2 Moderate increase 3 Marked increase
117	neutrophils_epith_03 Show the field ONLY if: [slides_number] = '3' or [slides_number] = '4' or [slides_number] = '5' or [slides_number] = '6' or [slides_number] = '7' or [slides_number] = '8'	Neutrophils in the epithelium Slide 3	radio 0 None 1 < 50% crypt involved 2 >50% crypt involved
118	ulceration_03 Show the field ONLY if: [slides_number] = '3' or [slides_number] = '4' or [slides_number] = '5' or [slides_number] = '6' or [slides_number] = '7' or [slides_number] = '8'	Ulceration Slide 3 *(defined as visible epithelial injury and regeneration and/or fibrin and neutrophils and/or tissue granulation)*	radio 0 Absent 1 Present
119	acute_inflam_03 Show the field ONLY if: [slides_number] = '3' or [slides_number] = '4' or [slides_number] = '5' or [slides_number] = '6' or [slides_number] = '7' or [slides_number] = '8'	Acute inflammatory cell infiltrate Slide 3	radio 0 None 1 Mild 2 Moderate 3 Severe
120	mucin_03 Show the field ONLY if: [slides_number] = '3' or [slides_number] = '4' or [slides_number] = '5' or [slides_number] = '6' or [slides_number] = '7' or [slides_number] = '8'	Mucin depletion Slide 3	radio 0 None 1 Mild 2 Moderate 3 Severe

121	neutrophils_lamina_03 Show the field ONLY if: [slides_number] = '3' or [slides_number] = '4' or [slides_number] = '5' or [slides_number] = '6' or [slides_number] = '7' or [slides_number] = '8'	Neutrophils in lamina propria Slide 3	radio 0 None 1 Mild 2 Moderate 3 Severe
122	basal_plasmacyt_03 Show the field ONLY if: [slides_number] = '3' or [slides_number] = '4' or [slides_number] = '5' or [slides_number] = '6' or [slides_number] = '7' or [slides_number] = '8'	Basal plasmacytosis Slide 3	radio 0 None 1 Mild 2 Moderate 3 Severe
123	serrated_abnor_03 Show the field ONLY if: [slides_number] = '3' or [slides_number] = '4' or [slides_number] = '5' or [slides_number] = '6' or [slides_number] = '7' or [slides_number] = '8'	Serrated architectural abnormalities Slide 3 *(defined as the presence of dilated crypts showing a scalloped lumen)*	radio 0 None 1 < 5% crypt involved 2 < 50% crypt involved 3 >50% crypt involved
124	notes_03 Show the field ONLY if: [slides_number] = '3' or [slides_number] = '4' or [slides_number] = '5' or [slides_number] = '6' or [slides_number] = '7' or [slides_number] = '8'	Notes Slide 3	notes
125	slide_04 Show the field ONLY if: [slides_number] = '4' or [slides_number] = '5' or [slides_number] = '6' or [slides_number] = '7' or [slides_number] = '8'	Slide 4	descriptive
126	code_04 Show the field ONLY if: [slides_number] = '4' or [slides_number] = '5' or [slides_number] = '6' or [slides_number] = '7' or [slides_number] = '8'	Biopsy code Slide 4	text
127	location_04 Show the field ONLY if: [slides_number] = '4' or [slides_number] = '5' or [slides_number] = '6' or [slides_number] = '7' or [slides_number] = '8'	Location Slide 4	dropdown 1 Rectum 2 Sigmoid 3 Descending colon 4 Transverse colon 5 Ascending colon 6 Cecum 7 Terminal ileum 8 Right colon 9 Left colon

128	nancy_score_04 Show the field ONLY if: [slides_number] = '4' or [slides_number] = '5' or [slides_number] = '6' or [slides_number] = '7' or [slides_number] = '8'	Nancy score Slide 4 *(defined as the quantity of lymphocytes and plasmacytes in the biopsy specimen)*	descriptive
129	inflam_infiltrate_04 Show the field ONLY if: [slides_number] = '4' or [slides_number] = '5' or [slides_number] = '6' or [slides_number] = '7' or [slides_number] = '8'	Chronic inflammatory infiltrate Slide 4 *(defined as the quantity of lymphocytes and plasmacytes in the biopsy specimen)*	radio 0 No increase 1 Mild but unequivocal increase 2 Moderate increase 3 Marked increase
130	neutrophils_epith_04 Show the field ONLY if: [slides_number] = '4' or [slides_number] = '5' or [slides_number] = '6' or [slides_number] = '7' or [slides_number] = '8'	Neutrophils in the epithelium Slide 4	radio 0 None 1 < 50% crypt involved 2 >50% crypt involved
131	ulceration_04 Show the field ONLY if: [slides_number] = '4' or [slides_number] = '5' or [slides_number] = '6' or [slides_number] = '7' or [slides_number] = '8'	Ulceration Slide 4 *(defined as visible epithelial injury and regeneration and/or fibrin and neutrophils and/or tissue granulation)*	radio 0 Absent 1 Present
132	acute_inflam_04 Show the field ONLY if: [slides_number] = '4' or [slides_number] = '5' or [slides_number] = '6' or [slides_number] = '7' or [slides_number] = '8'	Acute inflammatory cell infiltrate Slide 4	radio 0 None 1 Mild 2 Moderate 3 Severe
133	mucin_04 Show the field ONLY if: [slides_number] = '4' or [slides_number] = '5' or [slides_number] = '6' or [slides_number] = '7' or [slides_number] = '8'	Mucin depletion Slide 4	radio 0 None 1 Mild 2 Moderate 3 Severe
134	neutrophils_lamina_04 Show the field ONLY if: [slides_number] = '4' or [slides_number] = '5' or [slides_number] = '6' or [slides_number] = '7' or [slides_number] = '8'	Neutrophils in lamina propria Slide 4	radio 0 None 1 Mild 2 Moderate 3 Severe
135	basal_plasmacyt_04 Show the field ONLY if: [slides_number] = '4' or [slides_number] = '5' or [slides_number] = '6' or [slides_number] = '7' or [slides_number] = '8'	Basal plasmacytosis Slide 4	radio 0 None 1 Mild 2 Moderate 3 Severe
136	serrated_abnor_04 Show the field ONLY if: [slides_number] = '4' or [slides_number] = '5' or [slides_number] = '6' or [slides_number] = '7' or [slides_number] = '8'	Serrated architectural abnormalities Slide 4 *(defined as the presence of dilated crypts showing a scalloped lumen)*	radio 0 None 1 < 5% crypt involved 2 < 50% crypt involved 3 >50% crypt involved

137	notes_04 Show the field ONLY if: [slides_number] = '4' or [slides_number] = '5' or [slides_number] = '6' or [slides_number] = '7' or [slides_number] = '8'	Notes Slide 4	notes
138	slide_05 Show the field ONLY if: [slides_number] = '5' or [slides_number] = '6' or [slides_number] = '7' or [slides_number] = '8'	Slide 5	descriptive
139	code_05 Show the field ONLY if: [slides_number] = '5' or [slides_number] = '6' or [slides_number] = '7' or [slides_number] = '8'	Biopsy code Slide 5	text
140	location_05 Show the field ONLY if: [slides_number] = '5' or [slides_number] = '6' or [slides_number] = '7' or [slides_number] = '8'	Location Slide 5	dropdown 1 Rectum 2 Sigmoid 3 Descending colon 4 Transverse colon 5 Ascending colon 6 Cecum 7 Terminal ileum 8 Right colon 9 Left colon
141	nancy_score_05 Show the field ONLY if: [slides_number] = '5' or [slides_number] = '6' or [slides_number] = '7' or [slides_number] = '8'	Nancy score Slide 5 *(defined as the quantity of lymphocytes and plasmacytes in the biopsy specimen)*	descriptive
142	inflam_infiltrate_05 Show the field ONLY if: [slides_number] = '5' or [slides_number] = '6' or [slides_number] = '7' or [slides_number] = '8'	Chronic inflammatory infiltrate Slide 5 *(defined as the quantity of lymphocytes and plasmacytes in the biopsy specimen)*	radio 0 No increase 1 Mild but unequivocal increase 2 Moderate increase 3 Marked increase
143	neutrophils_epith_05 Show the field ONLY if: [slides_number] = '5' or [slides_number] = '6' or [slides_number] = '7' or [slides_number] = '8'	Neutrophils in the epithelium Slide 5	radio 0 None 1 < 50% crypt involved 2 >50% crypt involved
144	ulceration_05 Show the field ONLY if: [slides_number] = '5' or [slides_number] = '6' or [slides_number] = '7' or [slides_number] = '8'	Ulceration Slide 5 *(defined as visible epithelial injury and regeneration and/or fibrin and neutrophils and/or tissue granulation)*	radio 0 Absent 1 Present

145	acute_inflam_05 Show the field ONLY if: [slides_number] = '5' or [slides_number] = '6' or [slides_number] = '7' or [slides_number] = '8'	Acute inflammatory cell infiltrate Slide 5	radio 0 None 1 Mild 2 Moderate 3 Severe
146	mucin_05 Show the field ONLY if: [slides_number] = '5' or [slides_number] = '6' or [slides_number] = '7' or [slides_number] = '8'	Mucin depletion Slide 5	radio 0 None 1 Mild 2 Moderate 3 Severe
147	neutrophils_lamina_05 Show the field ONLY if: [slides_number] = '5' or [slides_number] = '6' or [slides_number] = '7' or [slides_number] = '8'	Neutrophils in lamina propria Slide 5	radio 0 None 1 Mild 2 Moderate 3 Severe
148	basal_plasmacyt_05 Show the field ONLY if: [slides_number] = '5' or [slides_number] = '6' or [slides_number] = '7' or [slides_number] = '8'	Basal plasmacytosis Slide 5	radio 0 None 1 Mild 2 Moderate 3 Severe
149	serrated_abnor_05 Show the field ONLY if: [slides_number] = '5' or [slides_number] = '6' or [slides_number] = '7' or [slides_number] = '8'	Serrated architectural abnormalities Slide 5 *(defined as the presence of dilated crypts showing a scalloped lumen)*	radio 0 None 1 < 5% crypt involved 2 < 50% crypt involved 3 >50% crypt involved
150	notes_05 Show the field ONLY if: [slides_number] = '5' or [slides_number] = '6' or [slides_number] = '7' or [slides_number] = '8'	Notes Slide 5	notes
151	slide_06 Show the field ONLY if: [slides_number] = '6' or [slides_number] = '7' or [slides_number] = '8'	Slide 6	descriptive
152	code_06 Show the field ONLY if: [slides_number] = '6' or [slides_number] = '7' or [slides_number] = '8'	Biopsy code Slide 6	text

153	location_06 Show the field ONLY if: [slides_number] = '6' or [slides_number] = '7' or [slides_number] = '8'	Location Slide 6	dropdown 1 Rectum 2 Sigmoid 3 Descending colon 4 Transverse colon 5 Ascending colon 6 Cecum 7 Terminal ileum 8 Right colon 9 Left colon
154	nancy_score_06 Show the field ONLY if: [slides_number] = '6' or [slides_number] = '7' or [slides_number] = '8'	Nancy score Slide 6 *(defined as the quantity of lymphocytes and plasmacytes in the biopsy specimen)*	descriptive
155	inflam_infiltrate_06 Show the field ONLY if: [slides_number] = '6' or [slides_number] = '7' or [slides_number] = '8'	Chronic inflammatory infiltrate Slide 6 *(defined as the quantity of lymphocytes and plasmacytes in the biopsy specimen)*	radio 0 No increase 1 Mild but unequivocal increase 2 Moderate increase 3 Marked increase
156	neutrophils_epith_06 Show the field ONLY if: [slides_number] = '6' or [slides_number] = '7' or [slides_number] = '8'	Neutrophils in the epithelium Slide 6	radio 0 None 1 < 50% crypt involved 2 >50% crypt involved
157	ulceration_06 Show the field ONLY if: [slides_number] = '6' or [slides_number] = '7' or [slides_number] = '8'	Ulceration Slide 6 *(defined as visible epithelial injury and regeneration and/or fibrin and neutrophils and/or tissue granulation)*	radio 0 Absent 1 Present
158	acute_inflam_06 Show the field ONLY if: [slides_number] = '6' or [slides_number] = '7' or [slides_number] = '8'	Acute inflammatory cell infiltrate Slide 6	radio 0 None 1 Mild 2 Moderate 3 Severe
159	mucin_06 Show the field ONLY if: [slides_number] = '6' or [slides_number] = '7' or [slides_number] = '8'	Mucin depletion Slide 6	radio 0 None 1 Mild 2 Moderate 3 Severe
160	neutrophils_lamina_06 Show the field ONLY if: [slides_number] = '6' or [slides_number] = '7' or [slides_number] = '8'	Neutrophils in lamina propria Slide 6	radio 0 None 1 Mild 2 Moderate 3 Severe

161	basal_plasmacyt_06 Show the field ONLY if: [slides_number] = '6' or [slides_number] = '7' or [slides_number] = '8'	Basal plasmacytosis Slide 6	radio 0 None 1 Mild 2 Moderate 3 Severe
162	serrated_abnor_06 Show the field ONLY if: [slides_number] = '6' or [slides_number] = '7' or [slides_number] = '8'	Serrated architectural abnormalities Slide 6 *(defined as the presence of dilated crypts showing a scalloped lumen)*	radio 0 None 1 < 5% crypt involved 2 < 50% crypt involved 3 >50% crypt involved
163	notes_06 Show the field ONLY if: [slides_number] = '6' or [slides_number] = '7' or [slides_number] = '8'	Notes Slide 6	notes
164	slide_07 Show the field ONLY if: [slides_number] = '7' or [slides_number] = '8'	Slide 7	descriptive
165	code_07 Show the field ONLY if: [slides_number] = '7' or [slides_number] = '8'	Biopsy code Slide 7	text
166	location_07 Show the field ONLY if: [slides_number] = '7' or [slides_number] = '8'	Location Slide 7	dropdown 1 Rectum 2 Sigmoid 3 Descending colon 4 Transverse colon 5 Ascending colon 6 Cecum 7 Terminal ileum 8 Right colon 9 Left colon
167	nancy_score_07 Show the field ONLY if: [slides_number] = '7' or [slides_number] = '8'	Nancy score Slide 7 *(defined as the quantity of lymphocytes and plasmacytes in the biopsy specimen)*	descriptive
168	inflam_infiltrate_07 Show the field ONLY if: [slides_number] = '7' or [slides_number] = '8'	Chronic inflammatory infiltrate Slide 7 *(defined as the quantity of lymphocytes and plasmacytes in the biopsy specimen)*	radio 0 No increase 1 Mild but unequivocal increase 2 Moderate increase 3 Marked increase
169	neutrophils_epith_07 Show the field ONLY if: [slides_number] = '7' or [slides_number] = '8'	Neutrophils in the epithelium Slide 7	radio 0 None 1 < 50% crypt involved 2 >50% crypt involved

170	ulceration_07 Show the field ONLY if: [slides_number] = '7' or [slides_number] = '8'	Ulceration Slide 7 *(defined as visible epithelial injury and regeneration and/or fibrin and neutrophils and/or tissue granulation)*	radio 0 Absent 1 Present
171	acute_inflam_07 Show the field ONLY if: [slides_number] = '7' or [slides_number] = '8'	Acute inflammatory cell infiltrate Slide 7	radio 0 None 1 Mild 2 Moderate 3 Severe
172	mucin_07 Show the field ONLY if: [slides_number] = '7' or [slides_number] = '8'	Mucin depletion Slide 7	radio 0 None 1 Mild 2 Moderate 3 Severe
173	neutrophils_lamina_07 Show the field ONLY if: [slides_number] = '7' or [slides_number] = '8'	Neutrophils in lamina propria Slide 7	radio 0 None 1 Mild 2 Moderate 3 Severe
174	basal_plasmacyt_07 Show the field ONLY if: [slides_number] = '7' or [slides_number] = '8'	Basal plasmacytosis Slide 7	radio 0 None 1 Mild 2 Moderate 3 Severe
175	serrated_abnor_07 Show the field ONLY if: [slides_number] = '7' or [slides_number] = '8'	Serrated architectural abnormalities Slide 7 *(defined as the presence of dilated crypts showing a scalloped lumen)*	radio 0 None 1 < 5% crypt involved 2 < 50% crypt involved 3 >50% crypt involved
176	notes_07 Show the field ONLY if: [slides_number] = '7' or [slides_number] = '8'	Notes Slide 7	notes
177	slide_08 Show the field ONLY if: [slides_number] = '8'	Slide 8	descriptive
178	code_08 Show the field ONLY if: [slides_number] = '8'	Biopsy code Slide 8	text

179	location_08 Show the field ONLY if: [slides_number] = '8'	Location Slide 8	dropdown 1 Rectum 2 Sigmoid 3 Descending colon 4 Transverse colon 5 Ascending colon 6 Cecum 7 Terminal ileum 8 Right colon 9 Left colon
180	nancy_score_08 Show the field ONLY if: [slides_number] = '8'	Nancy score Slide 8 *(defined as the quantity of lymphocytes and plasmacytes in the biopsy specimen)*	descriptive
181	inflam_infiltrate_08 Show the field ONLY if: [slides_number] = '8'	Chronic inflammatory infiltrate Slide 8 *(defined as the quantity of lymphocytes and plasmacytes in the biopsy specimen)*	radio 0 No increase 1 Mild but unequivocal increase 2 Moderate increase 3 Marked increase
182	neutrophils_epith_08 Show the field ONLY if: [slides_number] = '8'	Neutrophils in the epithelium Slide 8	radio 0 None 1 < 50% crypt involved 2 >50% crypt involved
183	ulceration_08 Show the field ONLY if: [slides_number] = '8'	Ulceration Slide 8 *(defined as visible epithelial injury and regeneration and/or fibrin and neutrophils and/or tissue granulation)*	radio 0 Absent 1 Present
184	acute_inflam_08 Show the field ONLY if: [slides_number] = '8'	Acute inflammatory cell infiltrate Slide 8	radio 0 None 1 Mild 2 Moderate 3 Severe
185	mucin_08 Show the field ONLY if: [slides_number] = '8'	Mucin depletion Slide 8	radio 0 None 1 Mild 2 Moderate 3 Severe
186	neutrophils_lamina_08 Show the field ONLY if: [slides_number] = '8'	Neutrophils in lamina propria Slide 8	radio 0 None 1 Mild 2 Moderate 3 Severe
187	basal_plasmacyt_08 Show the field ONLY if: [slides_number] = '8'	Basal plasmacytosis Slide 8	radio 0 None 1 Mild 2 Moderate 3 Severe

188	serrated_abnor_08 Show the field ONLY if: [slides_number] = '8'	Serrated architectural abnormalities Slide 8 *(defined as the presence of dilated crypts showing a scalloped lumen)*	radio 0 None 1 < 5% crypt involved 2 < 50% crypt involved 3 >50% crypt involved
189	notes_08 Show the field ONLY if: [slides_number] = '8'	Notes Slide 8	notes
190	pathology_assessment_complete	Section Header: *Form Status* Complete?	dropdown 0 Incomplete 1 Unverified 2 Complete