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

A COVID-19 na população portuguesa: uma análise de riscos competitivos

Margarida Teixeira Ribeiro

Mestrado em Bioestatística

Trabalho de Projeto orientado por:
Prof.ª Doutora Cristina Maria Tristão Simões Rocha

2022

Resumo

A COVID-19 é uma doença respiratória infeciosa causada pelo coronavírus SARS-CoV-2 que originou uma pandemia que se tem prolongado pelos anos de 2020 e 2021. Identificar os fatores de risco determinantes é fundamental para minimizar o risco de acontecimentos adversos, especialmente a morte.

Este estudo tem como objetivo realizar uma análise de sobrevivência recorrendo a modelos de riscos competitivos em indivíduos positivos à COVID-19, em Portugal, registados no Sistema Nacional de Vigilância Epidemiológica (SINAVE), com o propósito de analisar os fatores associados ao maior risco de ocorrência de óbito.

A variável dependente foi o tempo até a ocorrência de um dos acontecimentos possíveis, que são Morte por COVID-19, Recuperação e Morte por outra causa. As variáveis independentes foram sexo, grupo etário e Administração Regional de Saúde (ARS). Para a análise estatística foram utilizados vários métodos de inferência estatística, nomeadamente a estimativa da função de incidência cumulativa, teste de Gray e modelo de Fine e Gray.

No período abrangido pelo estudo, isto é, de 2 de março a 31 de dezembro de 2020 foram registadas 360 914 pessoas, das quais 5 197 tiveram como causa básica de morte COVID-19, 411 morreram por outra causa e 313 377 recuperaram. O tempo médio desde a infeção até a morte por COVID-19 foi 11,2 dias, até à morte por outra causa foi 13,5 dias e até à recuperação foi 16 dias. Para todos os grupos, definidos por sexo, grupo etário e ARS, a recuperação destaca-se por ser o acontecimento com maior probabilidade de ocorrência. Estima-se que indivíduos do sexo masculino têm um maior risco de morte por COVID-19 do que por outra causa, após infetados pelo vírus.

A partir da identificação do perfil dos pacientes com maior risco de óbito, devem ser delineadas estratégias de cuidados específicos para prevenir a evolução da doença que resulte num óbito.

Palavras-chave: análise de sobrevivência, COVID-19, fatores de risco, riscos competitivos

Abstract

COVID-19 is a contagious respiratory disease caused by the SARS-CoV-2 coronavirus, which has sparked a pandemic that will last into 2020 and 2021. Identifying decisive risk factors is critical to minimizing the risk of adverse events, especially death.

This study aimed to perform a survival analysis of Portuguese COVID-19-positive individuals registered in Sistema Nacional de Vigilância Epidemiológica (SINAVE) using a competing risk model, with the aim of analyzing the factors associated with the occurrence of death.

The dependent variable is the time until one of the possible outcomes occurs.The events of interest considered were Death from COVID-19, Recovery and Death from another cause. The independent variables were gender, age group and Regional Health Service. For statistical analysis, several statistical inference methods were used, namely estimation of cumulative correlation function, Gray's test, and Fine and Gray's model.

During the period covered by the study, from March 2 to December 31 in 2020, 360 914 people were registered, of whom 5 197 had COVID-19 as the underlying cause of death, 411 died from other causes, and 313 377 recovered. The median time from infection to death from COVID-19 was 11,2 days, death from other causes was 13,5 days, and recovery was 16 days. Recovery was the most likely event for all groups defined by sex, age group, and Regional Health Service. Men are estimated to have a higher risk of dying from COVID-19 after contracting the virus than from any other cause.

Based on the identification of the characteristics of patients at higher risk of death, specific care strategies should be developed to prevent the evolution of the disease leading to death.

Key-words: survival analysis, COVID-19, risk factors, competing risks

Agradecimentos

Desejo exprimir os meus agradecimentos a todos aqueles que, de alguma forma, permitiram que esta tese se concretizasse.

Em primeiro lugar gostaria de agradecer à Professora Doutora Cristina Simões Rocha, por ter aceite este desafio e acreditado neste trabalho, pela sua disponibilidade dispensada (mesmo fora do horário de expediente) e pelo encorajamento. O seu profundo conhecimento pelo tema e as suas sugestões foram fundamentais para a realização deste trabalho.

Deixo também um agradecimento muito especial à Direcção-Geral da Saúde pela disponibilização dos dados e pela oportunidade de estágio que me permitiu o contacto com conceitos e aplicações fundamentais da área. Em especial ao Dr. André Peralta-Santos, Dr. Pedro Casaca e Dra. Ana Sottomayor pela aprendizagem e ajuda, mas acima de tudo, pela amizade.

Aos meus colegas de mestrado, em particular ao Hugo que partilhou as mesmas desavenças durante o estágio e à Joana pelo companheirismo.

Aos meus colegas de trabalho, Henrique, João, Paulo e Sara, pela paciência, compreensão e motivação constante.

Agradeço aos meus amigos, Alexandra, Joana, Carolina, Patrícia, Daniela, Leonor, Beatriz, Nuno e Rita, que sempre me apoiaram e mostraram interesse pelo meu trabalho. Foram eles que sempre me ouviram e apoiaram, nos momentos de desânimo e de alento.

À Mara, amiga de longa data, que me acompanha desde o primeiro dia de aulas no ensino básico, que mostra sempre curiosidade e entusiasmo sobre o tema. Foi quem sempre me ajudou nos momentos mais difíceis e quem mais me apoiou nas minhas vitórias.

Ao meu Pedro Miguel, pela motivação e pelo incentivo diário. Pela amizade e amor e por todos os belos momentos. Por conseguir fazer-me sorrir, tanto nos bons momentos como nos momentos menos bons. Obrigada por todos os conselhos, pela ajuda em ultrapassar as dificuldades e a crescer como pessoa.

Por fim, aos meus pais, Zé Pedro e Alice, pelo apoio demonstrado, pelo incentivo, pela preocupação constante para que tudo acontecesse de modo a estar feliz. Ao Pedro Diogo e à Inês, pelo carinho e apoio na realização deste projecto. Por tudo, um grande bem haja.

Índice

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas

ARDS	Síndrome da insuficiência respiratória aguda
ARS	Administração Regional de Saúde
DGS	Direção-Geral da Saúde
DSIA	Direção de Serviços de Informação e Análise
KM	Kaplan-Meier
LVT	Lisboa e Vale do Tejo
OMS	Organização Mundial da Saúde
PCR	*Polimerase Chain Reaction*
RA	Região Autónoma
SARS-CoV-2	Coronavírus da síndrome respiratória aguda 2
SICO	Sistema de Informação dos Certificados de Óbito
SINAVE	Sistema Nacional de Vigilância Epidemiológica
SPMS	Serviços Partilhados do Ministério da Saúde

Capítulo 1

Introdução

1.1 A COVID-19

Em dezembro de 2019, na cidade de Wuhan na China foi reportado um surto de pneumonia de causa desconhecida que rapidamente se espalhou pelo país. O agente patogénico causador da doença é um novo coronavírus, que foi denominado coronavírus da síndrome respiratória aguda grave 2 ou SARS-CoV-2 (*severe acute respiratory syndrome coronavirus 2*) . A Organização Mundial da Saúde (OMS) designou por COVID-19 (*Coronavirus Disease*, 2019) a doença causada pelo novo vírus, tendo sido declarada uma emergência de saúde pública de importância internacional a 30 de janeiro de 2020 e a 11 de março de 2020 como uma pandemia. Desde então, tem havido um esforço global na produção de informação com a finalidade de analisar os aspetos clínicos e epidemiológicos e identificar os fatores de prognóstico da doença (Galvão & Roncalli, 2020).

A 2 de março de 2020, foram detetados em Portugal os dois primeiros casos de COVID-19 e passados 10 dias foi declarado estado de alerta no país.

No final de agosto de 2020, os dados da OMS mostravam que, a nível mundial, já existiam mais de 23 milhões de casos confirmados e 810 492 óbitos devidos à nova doença (WHO, 2020). Até à data de 7 de novembro 2021, Portugal registou 1 097 557 casos confirmados de infecção pelo SARS-CoV-2 e 18 203 óbitos por COVID-19.

A COVID-19 é uma doença infeciosa que afeta o sistema respiratório do ser humano e que pode dar origem a uma pneumonia severa. Quem contrai esta doença pode não apresentar sintomas, isto é, ser assintomático ou pode apresentar sintomas ligeiros ou graves. Os sintomas mais comuns associados a esta doença são, nomeadamente fadiga, tosse seca, febre e falta de ar. A COVID-19 propaga-se através da inalação de gotículas de saliva e de secreções respiratórias que podem ficar suspensas no ar quando uma pessoa infetada tosse ou espirra. Entre os indivíduos infetados, cerca de 20% necessitam de cuidados hospitalares que requerem vagas nos hospitais, profissionais de saúde e, em alguns casos, ventiladores disponíveis. As complicações desta doença podem levar a paragens respiratórias, síndrome da insuficiência respiratória aguda (ARDS), choque séptico, tromboembolismo e falência de múltiplos órgãos, incluindo coração, fígado ou rins. Estas complicações podem provocar a morte do paciente.

A melhor maneira de prevenir e retardar a transmissão é estar bem informado sobre a doença e como o vírus se espalha. Para as pessoas se protegerem devem manter a distância social, usar uma máscara adequada e lavar as mãos com frequência.

Desde o início da pandemia que a comunidade científica internacional reuniu esforços para o de-

senvolvimento de vacinas contra a COVID-19. De facto, o desenvolvimento e a rápida disponibilização a nível mundial de vacinas seguras e eficazes são elementos essenciais para o controlo da pandemia. A capacidade de uma vacina prevenir a doença grave, a hospitalização e a morte corresponde ao mais importante indicador de efetividade da vacinação contra a COVID-19, especialmente pela pressão imposta por esta doença nos sistemas de saúde. Por isso a vacinação tem vindo a desempenhar um papel central na preservação de vidas humanas, na contenção da pandemia, na proteção dos sistemas de saúde e no restabelecimento da economia e da vida social (DGS, 2020).

1.2 Vigilância da infeção pelo SARS-CoV-2

Durante a pandemia, a Direção-Geral da Saúde (DGS) tem elaborado um relatório semanal de vigilância da infeção pelo SARS-CoV-2. Nesse relatório analisam a situação epidemiológica, a ocupação dos serviços de saúde, a mortalidade e a situação internacional.

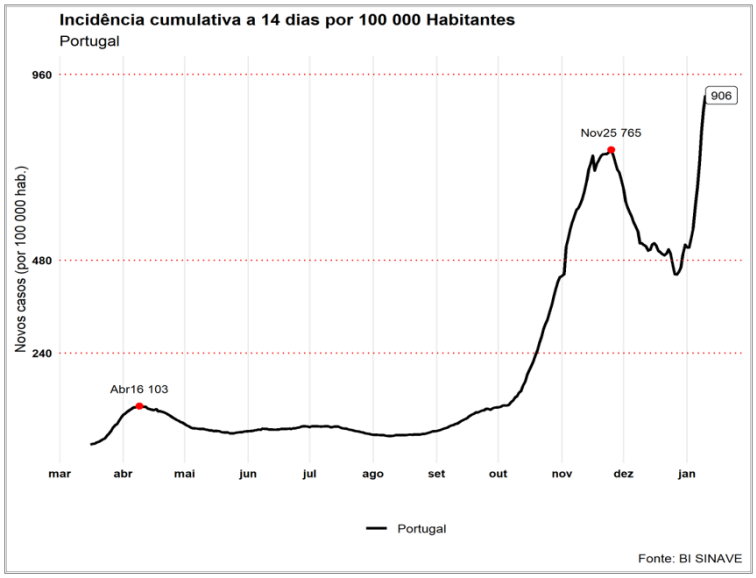
Para caracterizar a situação epidemiológica, compara-se o número de novos casos de infeção por SARS-CoV-2 da semana em estudo com as semanas anteriores. Para além dos novos casos, acompanhavam a evolução do número de casos recuperados, número de óbitos por COVID-19, número total de internados por COVID-19, número de internados por COVID-19 em Unidades de Cuidados Intensivos, número de amostras realizadas para testes laboratoriais e proporção de positividade.

SEMANA	07/12-13/12	14/12-20/12	21/12-27/12	28/12-03/01	04/01-10/01
Casos confirmados (N.º casos)	25 595	25 137	20 375	35 423	58 339
Casos recuperados (N.º casos)	28 434	25 326	22 079	22 737	27 637
Óbitos (N.º óbitos)	596	529	484	508	716
Internamento total (N.º casos no último dia do período em análise)	3 254	3 158	2 967	3 171	3983
Internamento em UCI (N.º casos no último dia do período em análise)	513	502	503	510	567
Amostras analisadas (N.º amostras laboratoriais)	220 466	241 974	232 815	227 157	336 001
Proporção de positividade (N.º amostras positivas sobre o total de amostras)	12,5%	11,1%	10,0%	16,0%	17,7%

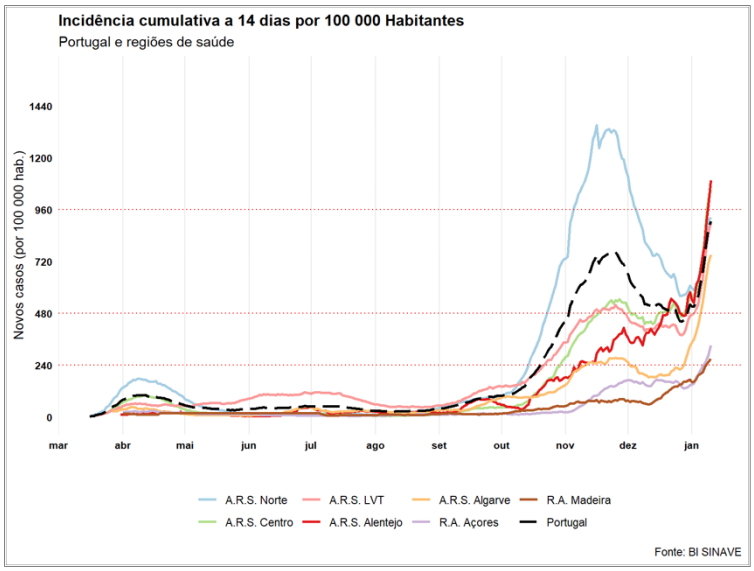
Fonte: BI SINAVE, SINAVE, Trace-Covid, SICO.

Figura 1.1: Evolução do número de casos de infeção por SARS-CoV-2/COVID-19 notificados, número de casos recuperados, número de óbitos COVID-19, número de internados por COVID-19 total, número de internados por COVID-19 em UCI, número de amostras realizadas para testes laboratoriais e proporção de positividade, em Portugal, por semana, entre 07/12/2020 e 10/01/2021. Fonte: DGS.

Outro aspeto muito importante para o estudo epidemiológico da COVID-19 consiste em analisar os valores da incidência cumulativa a 14 dias por 100 mil habitantes.



(a) Portugal



(b) ARS

Figura 1.2: Incidência cumulativa a 14 dias (por 100 000 habitantes), em Portugal e por região de saúde, de 03/03/2020 a 10/01/2021. Fonte: DGS

Estes gráficos representam o risco estimado de contrair a infeção por SARS-CoV-2 a nível nacional e regional. Como se pode observar, no ano de 2020 houve um pico de infeção a 16 de abril e a partir de setembro os casos aumentaram exponencialmente, atingindo um novo pico a 25 de novembro. Quanto ao gráfico (b), as Administrações Regionais de Saúde do Norte, do Centro e de Lisboa e Vale do Tejo (LVT) são as que apresentam os valores mais altos de incidência cumulativa a 14 dias por 100 mil habitantes. A nível global, foram considerados 5 níveis de risco. Uma incidência cumulativa a 14 dias por 100 mil habitantes até 120 é considerado baixo risco, de 120 até 240 um risco médio, de 240 até 480 um risco elevado, de 480 até 960 um risco extremo e, por fim, acima de 960 um risco extremamente elevado.

Esta análise foi feita a nível nacional e por Administração Regional de Saúde (ARS), mas também por concelho e por grupo etário.

Também foi recolhida informação sobre o número de surtos ativos, número de indivíduos sintomáticos e assintomáticos, número de profissionais de saúde infetados e se a infeção se deu fora do território nacional. Os novos casos de COVID-19, com infeção provocada por novas variantes também era reportada no relatório semanal.

No que diz respeito aos serviços de saúde, é importante perceber a evolução do número total de internamentos, número de internamentos em Unidades de Cuidados Intensivos e número de internamentos em enfermarias, assim como a distribuição dos internamentos por grupo etário e região.

Para avaliar o contágio da COVID-19, foi analisado o número de testes à COVID-19 e a percentagem de positividade ao vírus. Tendo em conta a elevada proporção de casos com notificações exclusivamente laboratoriais no SINAVE a nível nacional, foi realizada uma análise pelo tipo de teste. Os tipos existentes de testes são os Testes de Amplificação de Ácidos Nucleicos (TAAN, vulgo "testes de PCR") e Testes Rápidos de Antigénio (TRAg). Estes testes têm características diferentes, tais como a facilidade de acesso, notificação laboratorial e proporção de positividade.

Quanto à mortalidade, foi analisada a taxa de mortalidade por COVID-19, a 14 dias, por milhão de habitantes, a nível nacional e regional. Os dados utilizados para estudar a mortalidade provêm do Sistema de Informação dos Certificados de Óbito (SICO). A *baseline* utilizada para medir a gravidade da mortalidade provém do limiar do Centro Europeu de Prevenção e Controlo de Doenças (ECDC), que se situa nos 10%. Pela Figura 1.3 verifica-se que, Portugal encontrava-se acima deste limiar entre março e meados de junho de 2020. Em outubro o número de óbitos aumentou drasticamente, atingindo uma taxa de mortalidade acima dos 100 por milhão de habitantes no último mês de 2020.

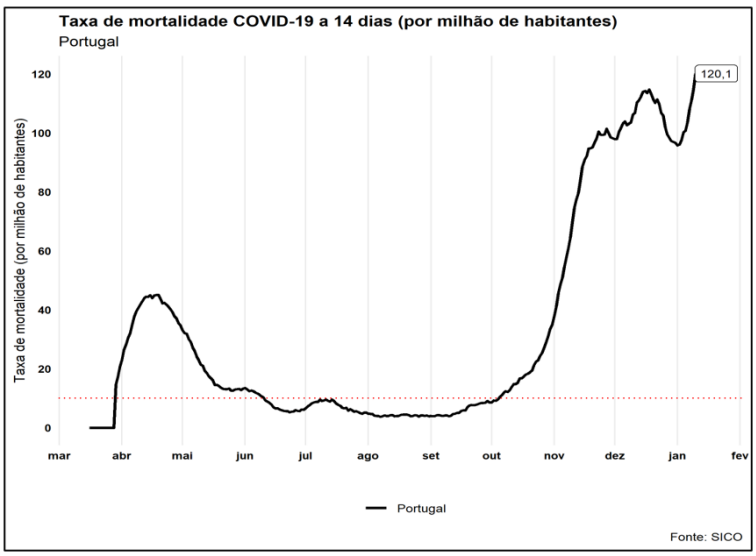


Figura 1.3: Taxa de mortalidade a 14 dias em Portugal entre 16 de março de 2020 e 10 de janeiro de 2021. Fonte: DGS.

Por fim, no relatório semanal da DGS era também reportada a situação epidemiológica nos países da União Europeia (UE), assim como a classificação de Portugal em relação aos outros países da UE quanto à incidência cumulativa de casos confirmados de COVID-19 a 14 dias por 100 000 habitantes.

1.3 Descrição dos sistemas de informação

Os casos confirmados de infeção por SARS-CoV-2 são contabilizados no Sistema Nacional de Vigilância Epidemiológica (SINAVE) através das notificações laboratoriais com resultado positivo, notificadas por laboratórios autorizados na plataforma SINAVE LAB, ou através das notificações clínicas com indicação de resultado positivo, notificadas por médicos na plataforma SINAVE MED (Figura1.4).

Estas notificações são distribuídas por área de jurisdição das Unidades de Saúde Pública (USP) no SINAVE. As Autoridades de Saúde e respetivas equipas de Saúde Pública de cada USP agregam as notificações relativas ao mesmo indivíduo, através do número de utente no SINAVE, criando um caso. Para cada caso, os dados das notificações são revistos e completados com o registo do inquérito epidemiológico realizado. Os casos são enviados pelo SINAVE para os Departamentos de Saúde Pública (DSP/DSPP) das Administrações Regionais de Saúde para serem validados e, posteriormente, para a Direção-Geral da Saúde, que comunica a informação aos decisores e organismos internacionais como o Centro Europeu de Prevenção e Controlo de Doenças ou a OMS, ao abrigo do Regulamento Sanitário Internacional (RSI).

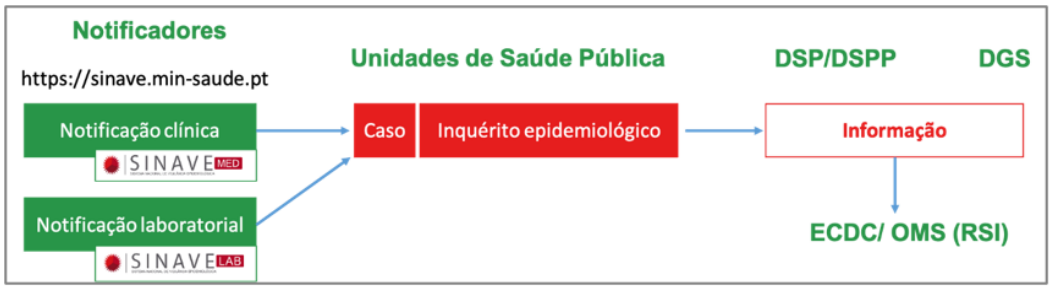


Figura 1.4: Esquema simplificado do fluxo de dados no SINAVE. Legenda: DSP/DSPP – Departamento de Saúde Pública/ Departamento de Saúde Pública e Planeamento; DGS - Direção-Geral da Saúde; ECDC/ OMS – European Centre for Disease Prevention and Control/ Organização Mundial da Saúde.

O grande aumento na incidência de casos de infeção pelo SARS-CoV-2, com início em setembro de 2020, obrigou as equipas de saúde locais a privilegiar a intervenção clínica e de Saúde Pública sobre o registo mais detalhado da informação e a criação de casos no SINAVE atempadamente. Assim, desde 16 de novembro de 2020, os casos passaram a ser contabilizados através do sistema BI SINAVE. Este sistema é uma solução de *Business Intelligence* que unifica todas as notificações laboratoriais, clínicas e inquéritos epidemiológicos realizados para o mesmo indivíduo através do número de utente, permitindo ter informação em tempo real. Neste sentido, existem casos para os quais no sistema há apenas informação proveniente de notificações laboratoriais, que incluem sexo, idade, local de residência e resultado do teste laboratorial. Assim, uma vez que nem todos os casos têm dados provenientes de notificações clínicas e inquéritos epidemiológicos, a informação apresentada deve ser interpretada cuidadosamente, em especial quando a proporção de casos apenas com notificações laboratoriais é elevada.

A recolha de dados relativos aos indivíduos que faleceram durante o tempo abrangido pelo estudo foi feita a partir do Sistema de Informação dos Certificados de Óbito (SICO). O SICO é o sistema de informação de mortalidade em Portugal que tem a finalidade de permitir a articulação das entidades envolvidas no processo de certificação dos óbitos, garantindo uma adequada utilização dos recursos, melhoria da qualidade, do rigor da informação e rapidez de acesso aos dados em condições de segurança

e no respeito pela privacidade dos cidadãos. A Direção-Geral da Saúde é a entidade responsável pela gestão e tratamento da base de dados do SICO e garante a vigilância epidemiológica da mortalidade identificando situações de risco para a saúde pública e a codificação das causas de morte de acordo com a Classificação Estatística Internacional de Doenças e problemas relacionados com a saúde - 10ª revisão (CID-10). A manutenção e o desenvolvimento da aplicação informática de suporte ao SICO são assegurados pelos Serviços Partilhados do Ministério da Saúde (SPMS).

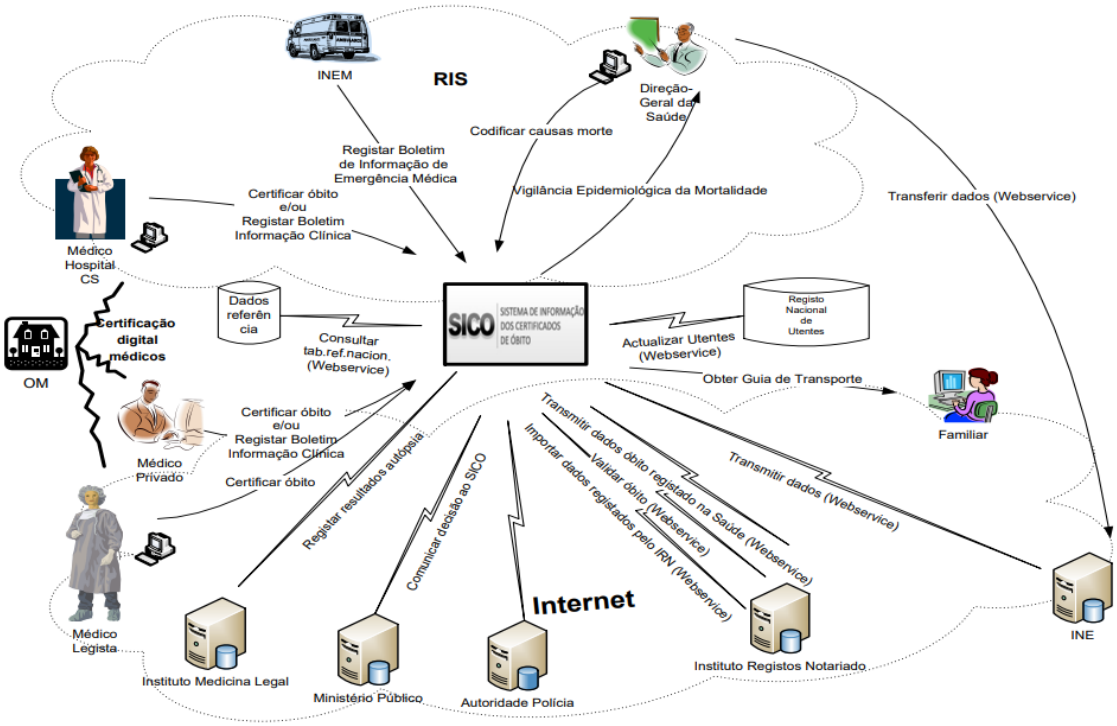


Figura 1.5: Arquitetura do Sistema de Informação dos Certificados de Óbito. Fonte:https://www.dgs.pt/ficheiros-de-upload-2013/sico-procedimentos-pdf.aspx

Com o aparecimento da COVID-19 no mundo, foi importante tentar determinar a letalidade desta nova doença e para isso era importante haver uma uniformização do uso da codificação básica quando a causa de morte era por COVID-19. A 20 de abril de 2020, a Organização Mundial de Saúde, lançou o documento Diretrizes Internacionais para a Certificação e Classificação (Codificação) da COVID-19 como Causa de Morte, onde indica vários casos em que a causa básica deve ser considerada COVID-19. Neste documento também é referido que utentes com comorbilidades, ou seja, indivíduos com condições crónicas existentes ou com sistema imunitário comprometido devido a debilidade, apresentam uma evidência crescente que estão em maior risco de morte devido à COVID-19.

De forma a existir um acompanhamento dos indivíduos infetados com SARS-CoV-2 desde o dia de positividade à COVID-19 até à sua morte, utilizou-se o número de utente para se fazer o cruzamento entre estes dois sistemas de informação, o SINAVE e o SICO.

É de frisar que para outras doenças, a recuperação não é um acontecimento registado pelo SINAVE. A COVID-19 mostrou-se singular em muitas vertentes, sendo a notificação de recuperação uma delas e esta foi feita a pedido da União Europeia, sendo um registo excecional.

Devido a constrangimentos, as informações relativas à recuperação dos indivíduos tratam-se de uma

estimativa de tempo médio da recuperação por grupo etário e segundo o período em que se deu a infeção por SARS-CoV-2. Por isso, os dados relativos aos recuperados são de baixa qualidade.

1.4 Motivação

A maior parte dos estudos desenvolvidos até ao final de 2020 relativos à COVID-19, é proveniente de cenários internacionais e de contexto clínico-hospitalar.

Em seguida, são referidos alguns desses estudos que utilizam métodos de Análise de Sobrevivência.

Um estudo que analisou a gravidade e a mortalidade da COVID-19 em adultos em Wuhan, incluiu todos os indivíduos com COVID-19 que foram admitidos no *Tongji Hospital* entre 26 de janeiro e 5 de fevereiro de 2020 (Li et al., 2020). Dos 548 pacientes, registaram o género, a idade, Índice de Massa Corporal (IMC), como se deu o contágio da doença, hábitos tabágicos, comorbilidades, medicação habitual, sintomas, a gravidade da doença (severa ou não severa) e o tipo de teste realizado (Teste PCR ou Teste Rápido de Antigénio). Foram também registadas as complicações que surgiram durante a hospitalização dos pacientes. A observação dos pacientes terminou quando ocorria a morte ou à data final do estudo. Do modelo de regressão de Cox de riscos proporcionais ajustado, concluiu-se que indivíduos do sexo masculino, de idade mais avançada, com leucocitose, com elevado nível de desidrogenase lática (DHL) e com hiperglicemia apresentavam um maior risco de morte ou de manifestação severa da COVID-19.

Noutro estudo, cujos dados provinham do Sistema de Vigilância Epidemiológica para Doenças Respiratórias Virais do Ministério da Saúde do México (*Epidemiological Surveillance System for Viral Respiratory Diseases of the Mexican Ministry of Health*), investigou-se o impacto da COVID-19 na população mexicana (Salinas-Escudero et al., 2020). O principal objetivo do estudo era identificar os fatores de risco associados à morte por COVID-19. Para esse fim, os autores recorreram a métodos clássicos da análise de sobrevivência como o estimador de Kaplan-Meier e o modelo de regressão de Cox. A informação registada de cada um dos 16 752 indivíduos foi o sexo, idade, nacionalidade, local de residência, estado migratório, doenças crónicas, imunossupressão, hábitos tabágicos e gravidez. Pelo modelo de Cox ajustado, concluiu-se que os indivíduos de sexo masculino com mais de 75 anos apresentavam um maior risco de morrer por COVID-19.

Um outro estudo pretendeu analisar a mortalidade dos casos de COVID-19 na Colômbia, entre março e julho de 2020 (Malagón-Rojas et al., 2021). Com base em dados do Instituto Nacional da Saúde, foi realizada uma análise de sobrevivência em que a data de início de observação dos indivíduos é a data de aparecimento de sintomas e a data final corresponde à data de óbito ou final do estudo. Neste estudo participaram 118 738 indivíduos. Os fatores de risco foram determinados pelo modelo de regressão de Cox. Concluiu-se que ser do sexo masculino, pertencer a uma faixa etária mais envelhecida e estar hospitalizado, está associado um maior risco de morte por COVID-19.

Um estudo realizado no estado do Rio Grande, Brasil, utilizou métodos de análise de sobrevivência em 52 607 indivíduos com COVID-19 registados nos Sistemas de Informação em Saúde, com o propósito de analisar os fatores associados ao risco de morte por COVID-19 (Galvão & Roncalli, 2020). Deste estudo concluíu-se que indivíduos do sexo masculino, com idade idade igual ou superior a 80 anos, com

uma cor de pele não branca e que possuam comorbilidades têm um maior risco de morrer por COVID-19.

De um modo geral, os estudos sugerem que casos mais graves da doença que evoluem para pneumonia são, com maior probabilidade, pacientes mais velhos, do sexo masculino e com comorbilidades. Contudo constatou-se a necessidade de mais estudos para esclarecer as características epidemiológicas da COVID-19, bem como identificar os fatores de risco e de prognóstico para a morte por COVID-19.

Assim sendo, é de realçar a importância de realizar este tipo de estudos em Portugal, utilizando para tal os dados produzidos pelos sistemas de informação SINAVE e SICO. Desta forma, o presente estudo tem o objetivo de realizar uma análise de sobrevivência de dados de indivíduos diagnosticados com COVID-19 registados no SINAVE e identificar os fatores com influência no risco de morte por COVID-19.

1.5 Descrição do estudo

Este trabalho descreve uma análise de dados de sobrevivência na presença de riscos competitivos em indivíduos notificados como casos de COVID-19, em Portugal desde o dia 2 de março até ao dia 31 de dezembro de 2020. Para aqueles que faleceram nesse período, é adicionada a informação registada no seu certificado de óbito como a data de óbito e a causa básica de morte. Os dados disponibilizados tiveram de passar por muitos processos de limpeza e de anonimização que limitaram e influenciaram a qualidade de certas variáveis.

Como referido anteriormente, devido à grande proporção de casos que têm apenas notificações laboratoriais, as variáveis que apresentam uma maior qualidade são as variáveis sexo, idade e Administração de Saúde à qual o paciente pertence.

A variável resposta representa o tempo (em dias) até a ocorrência do primeiro acontecimento (morte por COVID-19, morte por outra causa ou recuperação) a partir da data em que o teste realizado foi positivo para SARS-CoV-2. A data de fim de observação corresponde à data do óbito, ou à data em que se deu a recuperação ou à data final do estudo.

O objetivo deste estudo visa identificar quais os fatores que têm uma influência significativa no tempo de vida desde o dia de positividade à COVID-19 até à morte devido a esta doença.

1.6 Definição das variáveis

As variáveis de maior relevância para este estudo são apresentadas na Tabela 1.1. A variável de identificação é a `Id_Utente`, que contém um número de identificação fictício de cada utente. A variável `Estado` foi criada para identificar o primeiro acontecimento que o utente sofreu, sendo que se nenhum acontecimento ocorreu até ao fim do estudo, o utente foi codificado como Caso Ativo, que corresponde a uma observação censurada. A variável `Vaga` identifica em que período de tempo se deu a infeção por SARS-CoV-2. As restantes variáveis categóricas são: `sexo` (masculino, feminino), a variável `grupo etário` (até aos 20 anos, dos 20 aos 60 anos, dos 60 aos 70 anos, dos 70 até aos 80 anos e mais de 80 anos) e `ARS` (ARS Norte, ARS Centro, ARS Lisboa e Vale do Tejo, ARS Alentejo, ARS Algarve, RA Madeira e RA Açores).

Variável	Descrição	Tipo	Categoria
Id_Utente	Identificador do utente	Nominal	
Data_confirmado	Data de positividade à COVID-19		
Data_do_óbito	Data de ocorrência do óbito		
Data_recuperação	Data em que o utente é dado como recuperado		
Estado	Desfecho do utente	Categórica	1-Recuperação; 2-Morte por COVID-19; 3-Morte por outra causa; 4-Caso ativo
Tempo	Tempo entre a data de positividade e o desfecho	Numérica contínua	
Idade_cat	Grupo etário a que o utente pertence	Categórica	1-[00;20] 2-]20;60] 3-]60;70] 4-]70;80] 5- $>$ 80
Sexo	Sexo do utente	Categórica	F-Feminino M-Masculino
Vaga	Período em que o utente foi infetado	Categórica	1ªVaga - 03/mar-31/ago 2ªVaga- 01/set-31/nov 3ªVaga- 01/dez-31/dez
ARS	Região de saúde do utente	Categórica	ARS Norte ARS Centro ARS LVT ARS Alentejo ARS Algarve RA Madeira RA Açores

Tabela 1.1: Variáveis consideradas neste estudo

1.7 Critérios de inclusão e exclusão

Neste estudo estão incluídas todas as pessoas notificadas como casos positivos à COVID-19, que realizaram um teste e que este foi registado no Sistema Nacional de Vigilância Epidemiológica (SINAVE). Por sua vez, todos os indivíduos que recorreram aos testes de antigénio realizados em farmácias e aos auto-testes que não tinham obrigatoriedade de notificação ao sistema, não fazem parte do estudo bem como os indivíduos que não apresentavam indicação de sexo, idade ou administração regional de saúde.

A DGS trata as observações que não apresentam idade/sexo ou que apresentam idades muito avançadas, como erros de inserção por parte dos profissionais de saúde ou testes ao sistema, excluindo-os das análises e reportando às Administrações Regionais de Saúde (ARS) para averiguarem se são realmente testes ao sistema, falta de informação do utente ou erros. Por isto e pelo facto dos diferentes estudos (Li et al., 2020; Salinas-Escudero et al., 2020) mostrarem que o sexo e a idade são fatores com influência na mortalidade, foram excluídos do estudo todas as observações sem indicação do sexo e idade ou observações com uma idade igual ou superior a 110 anos. Como é de igual interesse averiguar se a região de saúde onde o indivíduo reside é relevante para a ocorrência dos acontecimentos em estudo, todas as pessoas que não apresentavam a região de saúde à qual a sua residência pertencia, não foram incluídas.

Capítulo 2

Análise de Sobrevivência

2.1 Introdução

A Análise de Sobrevivência tem como objetivo estudar o tempo decorrido desde um instante inicial bem definido até à ocorrência de um certo acontecimento. O acontecimento pode ser a morte, recaída de certa doença, alta hospitalar, arranjar um emprego, ou qualquer outro acontecimento de interesse.

A Análise de Sobrevivência engloba um conjunto de modelos e métodos estatísticos que se aplicam em áreas muito distintas como economia, física, engenharia, sociologia, psicologia, demografia e ciência biomédicas. Apesar do acontecimento de interesse poder assumir formas muito diversas, é habitualmente usado o termo ”morte”para designar a ocorrência do acontecimento. Tempo de vida designa o tempo decorrido desde o instante inicial bem definido até à ocorrência de um acontecimento de interesse. Os dados de sobrevivência representam o tempo de vida ou tempo de sobrevivência de indivíduos que pertencem a uma determinada população.

Um aspecto importante a considerar é a existência de dados censurados, que surgem quando, para alguns indivíduos em estudo, não é observada a realização do acontecimento de interesse durante o período em que esses indivíduos estão em observação. A Análise de Sobrevivência permite realizar a análise estatística de dados que representam tempos em observação que realmente correspondem ao tempo até ao instante em que se deu o acontecimento, juntamente com os dados censurados, não havendo desta forma perda de informação.

Habitualmente, para cada indivíduo em estudo, são recolhidos valores de outras variáveis, designadas por variáveis explicativas ou covariáveis, que representam fatores que podem influenciar o tempo de sobrevivência. Assim, a análise de regressão é também nesta área uma ferramenta estatística muito útil e, por conseguinte, foram desenvolvidos modelos de regressão adequados às especificidades dos dados de sobrevivência.

Neste capítulo, salvo indicações em contrário, a referência bibliográfica será Rocha & Papoila (2009).

2.2 Conceitos básicos

Seja T uma variável aleatória não negativa contínua, que representa o tempo de vida de indivíduos pertencentes a uma dada população homogénea, i. e., em que os indivíduos não diferem entre si relativamente a fatores suscetíveis de influenciar a sua sobrevivência.

Função de sobrevivência

Define-se função de sobrevivência no instante t como sendo a probabilidade de um indivíduo sobreviver para além do instante t. Trata-se de uma função monótona decrescente e contínua.

$$S(t) = P(T > t), \quad t \geq 0. \tag{2.1}$$

Por consequência, a função densidade de probabilidade é dada por $f(t) = -S'(t)$.

Função de risco

A distribuição de T pode também ser caracterizada pela função de risco que especifica a taxa instantânea de morte no instante t, condicional à sobrevivência do indivíduo até esse instante:

$$h(t) = \lim_{dt \to 0^+} \frac{P(t < T \leq t + dt | T \geq t)}{dt}. \tag{2.2}$$

A função de risco verifica as seguintes propriedades:

1. $h(t) \geq 0$
2. $\int_0^\infty h(t)dt = \infty$

Pela Equação (2.2), é possível obter algumas relações entre a função de sobrevivência, a função de densidade de probabilidade e a função de risco:

$$h(t) = \frac{f(t)}{S(t)} = -\frac{S^{'}(t)}{S(t)} = -\frac{d \log S(t)}{dt}. \tag{2.3}$$

Como $S(0) = 1$,

$$S(t) = \exp\left(-\int_0^t h(u)\, du\right). \tag{2.4}$$

A função de risco descreve a evolução da probabilidade instantânea de morte de um indivíduo, ao longo do tempo. Representa um aspeto da distribuição do tempo de vida que tem significado físico

imediato.

As formas mais comuns da função de risco são:

- **Monótona crescente:** surge em situações em que os indivíduos são observados num período da sua vida durante o qual ocorre um envelhecimento gradual, isto é, a proporção de indivíduos que morrem num dado instante, entre os sobreviventes nesse instante, aumenta com o tempo.
- **Monótona decrescente:** acontece quanto mais tempo um indivíduo sobrevive, menor é a probabilidade de morte no instante subsequente. Esta é uma das formas menos comuns.
- **Constante:** caracteriza, univocamente, a distribuição do tempo de vida como exponencial.
- ***Bathtub-shaped*:** esta forma acontece em populações em que os indivíduos são seguidos desde o nascimento até à morte real, sejam populações de seres vivos ou de objetos manufaturados.
- ***Hump-shaped* / unimodal:** o risco de morte é inicialmente crescente e decresce ao fim de algum tempo.

Função de risco cumulativa

Esta função não negativa crescente é definida por:

$$H(t) = \int_0^t h(u)\,du, \quad t \geq 0. \tag{2.5}$$

Logo, (2.4) é equivalente a

$$S(t) = \exp\{-H(t)\}. \tag{2.6}$$

Censura

Um aspeto importante a considerar na análise de dados de sobrevivência é a possibilidade de existência de dados censurados, que ocorrem quando para alguns indivíduos em estudo, não é observada a realização do acontecimento de interesse durante o período em que esses indivíduos estão sob observação.

Pode-se dizer que se dispõe apenas de informação parcial sobre o tempo de vida desses indivíduos, mas o período de tempo em observação deve ser registado, para que não haja perda de informação.

Existem vários tipos de censura:

1. **Censura à direita:** sucede quando o tempo em observação do indivíduo termina antes da ocorrência do acontecimento de interesse, apenas se sabe que o seu tempo de vida excede esse valor. Um dos tipos de censura à direita mais frequente é a censura aleatória. Este tipo surge quando os indivíduos entram no estudo de forma aleatória. Se o estudo terminar numa data pré-fixada, então o tempo decorrido desde que um indivíduo entra em estudo até ao final deste é aleatório.
2. **Censura à esquerda:** acontece quando apenas se sabe que o tempo de vida é menor do que o tempo que foi registado.

3. **Censura intervalar:** ocorre quando não é possível observar o instante exato em que ocorre o acontecimento de interesse mas apenas se sabe que tenha ocorrido num certo intervalo aleatório de tempo.

O mecanismo de censura diz-se independente ou não informativo quando os indivíduos que são censurados no instante t são representativos de todos os indivíduos, com as mesmas características, que sobreviveram até t. Em qualquer instante, os indivíduos não podem ser censurados por apresentarem um risco de morte muito elevado ou baixo.

Neste trabalho, o único tipo de censura existente é a censura à direita.

Truncatura

A truncatura surge quando, devido a um processo de seleção inerente ao planeamento do estudo, apenas são estudados os indivíduos a que ocorreu determinado acontecimento. A truncatura consiste numa condição que "oculta"certos indivíduos de modo que o investigador não se apercebe da sua existência.

Neste trabalho não existe qualquer tipo de truncatura.

2.3 Estimação não paramétrica da função de sobrevivência

Como referido anteriormente, a função de sobrevivência $S(t)$, representa a probabilidade de tempo de sobrevivência ser superior a um instante t.

Estimador de Kaplan-Meier

Quando não existe censura, a função de sobrevivência num dado instante t é estimada pela proporção de indivíduos que sobreviveram para além do instante t, ou seja, a proporção de tempos de vida observados de valor superior a t.

Kaplan e Meier (1958) propuseram um estimador não paramétrico da função de sobrevivência, quando existem observações censuradas, que é designado por estimador de Kaplan-Meier. O estimador de Kaplan-Meier é o estimador mais usado em estudos clínicos.

Sejam $t_{(1)},...,t_{(k)}$ os instantes de morte distintos numa amostra de dimensão n $(k \leq n)$, d_i o número de mortes ocorridas em $t_{(i)}$ e n_i o número de indivíduos em risco em $t_{(i)}$. O estimador de Kaplan-Meier da função de sobrevivência é dado por

$$\hat{S}(t) = \prod_{i:t_{(i)} \leq t} \left(1 - \frac{d_i}{n_i}\right), \tag{2.7}$$

sendo $\hat{S}(t) = 1$ para $0 \leq t < t_{(1)}$. Quando um instante de morte e um instante de censura são registados com o mesmo valor, considera-se que o instante de morte precede o instante de censura.

A estimativa da variância de $\hat{S}(t)$ é dada pela seguinte expressão, conhecida por fórmula de Greenwood:

$$\widehat{\text{Var}}[\hat{S}(t)] = [\hat{S}(t)]^2 \sum_{i:t_{(i)} \leq t} \frac{d_i}{n_i(n_i - d_i)}. \tag{2.8}$$

O estimador de Kaplan-Meier tem as seguintes propriedades:

- $\hat{S}(t) = 0$ para $t \geq t_{(k)}$, se $t_{(k)}$ for a maior observação registada, isto é, se a maior observação for não censurada;
- Se a maior observação registada t^* for censurada, então $\hat{S}(t)$ nunca toma o valor zero e considera-se que a estimativa está definida apenas até esse instante, sendo $\hat{S}(t) = \hat{S}(t_{(k)})$ para $t_{(k)} \leq t \leq t^*$;
- $\hat{S}(t)$ é uma função em escada, com saltos nos instantes de morte observados;
- $\hat{S}(t)$ é um estimador consistente de $S(t)$ e, sob certas condições de regularidade, pode ser considerado como um estimador de máxima verosimilhança não paramétrico de $S(t)$.

2.4 Estimação não paramétrica da função de risco cumulativa

A função $H(t)$ pode ser estimada por

$$\hat{H}(t) = -\log \hat{S}(t), \tag{2.9}$$

onde $\hat{S}(t)$ é o estimador de Kaplan-Meier da função de sobrevivência. Um estimador alternativo e mais usual, com melhor comportamento para pequenas amostras, é o estimador de Nelson-Aalen definido por

$$\tilde{H}(t) = \sum_{i:t_{(i)} \leq t} \frac{d_i}{n_i}.$$

É também designado por função de risco cumulativa empírica. A estimativa da variância de $\tilde{H}(t)$ é dada por

$$\widehat{\text{Var}}\left\{\tilde{H}(t)\right\} = \sum_{i:t_{(i)} \leq t} \frac{d_i}{n_i^2}. \tag{2.10}$$

A função de risco cumulativa é importante na identificação de modelos para o tempo de vida com base no comportamento da função de risco, visto que $H(t)$ é uma função não decrescente que será linear se $h(t)$ for constante, convexa se $h(t)$ for crescente ou côncava se $h(t)$ for decrescente.

Além disso, como o declive da função de risco cumulativa fornece informação acerca da forma da função de risco, é possível obter uma estimativa grosseira desta função a partir do declive do gráfico do

estimador de Nelson-Aalen. Dado que a estimativa assim obtida é frequentemente difícil de interpretar, é então necessário recorrer a métodos de suavização.

Note-se que

$$\hat{H}(t) = -\sum_{i:t_{(i)}\leq t} \log\left(1-\frac{d_i}{n_i}\right) = \sum_{i:t_{(i)}\leq t}\left(\frac{d_i}{n_i}+\frac{d_i^2}{2n_i^2}+...\right).$$

Verifica-se que o estimador de Nelson-Aalen pode ser considerado como uma aproximação de 1ª ordem do estimador de Kaplan-Meier da função de risco cumulativa. Para modelos contínuos os dois estimadores são assintoticamente equivalentes e dão resultados que não diferem muito, excepto quando há poucos indivíduos em risco.

Sendo o estimador de Nelson-Aalen um estimador da função de risco cumulativa, é óbvio que a partir dele se pode obter um estimador da função de sobrevivência, também designado por estimador de Breslow, dado por

$$\tilde{S}(t) = \exp(-\tilde{H}(t)) = \prod_{i:t(i)\leq t} \exp\left(-\frac{d_i}{n_i}\right).$$

Visto que $e^{-x} \approx 1-x$ quando x é pequeno, tem-se que $\exp(-d_i/n_i) \approx 1-d_i/n_i$ enquanto n_i for grande comparativamente a d_i, isto é, d_i/n_i pequeno. Como regra prática refere-se, por vezes, $n_i \geq 10d_i$. Em qualquer instante t, a estimativa de Nelson-Aalen da função de sobrevivência é sempre superior ou igual à estimativa de Kaplan-Meier, visto que $\exp(-x) \geq 1-x$; logo $\hat{S}(t) \leq \tilde{S}(t), \forall t > 0$. É de salientar que, se a maior observação registada t_{max} for não censurada, para $t \geq t_{max}, \hat{S}(t) = 0$ enquanto que $\tilde{S}(t) > 0$. Embora, para pequenas amostras, o estimador de Nelson-Aalen apresente um melhor comportamento do que o estimador de Kaplan-Meier, em muitas situações as estimativas serão muito semelhantes.

2.5 Alguns testes não paramétricos

Para comparar a distribuição do tempo de vida para vários grupos de indivíduos, o primeiro método é obter a estimativa de Kaplan-Meier da função de sobrevivência para cada grupo e, seguidamente, representar graficamente as curvas de sobrevivência permitindo ter uma ideia do seu comportamento e avaliar, de um modo informal, se existem diferenças entre os vários grupos, quanto ao seu padrão de sobrevivência. Também é recomendado representar graficamente as estimativas de Nelson-Aalen da função de risco cumulativa para os vários grupos, obtendo informação sobre a evolução do risco. Contudo, para uma análise mais rigorosa da existência de diferenças significativas entre os vários grupos, é necessário recorrer a testes de hipóteses.

No que se segue, onde será explicado alguns testes de hipóteses para a comparação das curvas de sobrevivência, vai-se considerar dois grupos de indivíduos, em que $S_i(t)$ representa a função de sobrevivência de um indivíduo no i-ésimo grupo, sendo então as hipóteses a testar

$$H_0 : S_1(t) = S_2(t) \qquad vs \qquad H_1 : S_1(t) \neq S_2(t).$$

Teste log-rank

Sejam m e n as dimensões das amostras correspondentes aos grupos 1 e 2, respectivamente e seja $t_1 < ... < t_k$ os instantes distintos de ocorrência do acontecimento de interesse relativos aos $m+n$ pacientes. Onde d_j representa o número de acontecimentos ocorridos em $t_j, j = 1,...,k$, d_{ij} o número de acontecimentos ocorridos em t_j no grupo $i, i = 1,2, n_j$ o número de indivíduos em risco imediatamente antes de $t_j, j = 1,...,k$ e n_{ij}.

A informação relevante em cada instante t_j encontra-se resumida na tabela seguinte.

Grupo	**No. de mortes em t_j**	**No. de sobrev. para além de t_j**	**No. de ind. em risco em t_j**
1	d_{1j}	$n_{1j} - d_{1j}$	n_{1j}
2	d_{2j}	$n_{2j} - d_{2j}$	n_{2j}
	d_j	$n_j - d_j$	n_j

Mantel e Haenzel (1959) sugeriram considerar a distribuição condicional das frequências observadas em cada célula, dados os totais marginais, sob a validade da hipótese nula. Isto implica considerar a distribuição da frequência de apenas uma célula, d_{ij}, visto que as outras frequências ficam implicitamente determinadas pelos totais marginais fixos. Então, supondo H_0 verdadeira, a distribuição de d_{1j}, condicional aos valores marginais, é hipergeométrica

$$p(d_{1j}|d_j, n_j) = \frac{\binom{d_j}{d_{1j}}\binom{n_j - d_j}{n_{1j} - d_{1j}}}{\binom{n_j}{n_{1j}}}.$$

Sob H_0, o valor médio e a variância condicionais de d_{1j} são, respectivamente,

$$e_{1j} = n_{1j}\frac{d_j}{n_j}$$

e

$$v_{1j} = \frac{n_{1j}n_{2j}d_j(n_j - d_j)}{n_j^2(n_j - 1)}.$$

Note-se que e_{1j} é o número esperado de acontecimentos ocorridos no instante t_j no grupo 1, sob H_0. De facto, d_j/n_j é a probabilidade de ocorrência do acontecimento em t_j no grupo $i, i = 1,2$, condicional à sobrevivência até esse instante, visto que, sob H_0, esta probabilidade é igual nos dois grupos.

Para combinar a informação contida nas k tabelas de contingência, de modo a obter uma medida global do desvio dos valores observados de d_{1j} em relação aos valores esperados, consideremos a estatística

$$U = \sum_{j=1}^{k} (d_{1j} - e_{1j}).$$

Então, $E(U) = 0$ e

$$var(U) = \sum_{j=1}^{k} v_{1j}.$$

A estatística de teste proposta por Mantel e Haenszel é

$$Q = \frac{U^2}{var(U)}$$

que tem distribuição assintótica χ^2_1, sob H_0.

O teste log-rank é o mais potente na deteção de afastamentos da hipótese de igualdade das distribuições de sejam do tipo de riscos proporcionais. É ainda bastante potente para alternativas em que as funções de risco sejam não proporcionais mas não se cruzem. Quando as funções de risco se cruzam, o teste log-rank pode não conseguir detetar diferenças significativas entre as curvas de sobrevivência.

Teste de Gehan

Este teste trata-se de uma generalização do teste de Mann-Whitney-Wilcoxon para dados censurados, proposto em 1965 por Gehan, como o nome sugere. É também designado por teste de Wilcoxon generalizado. Forme-se então a amostra conjunta de tempos de observação de dimensão $m+n$.

Seja

$$U_G = \sum_{j=1}^{k} n_j (d_{1j} - e_{1j})$$

onde $e_{1j} = n_{1j} d_j / n_j$. A variância da estatística U_G é dada por $V_G = \sum_{j=1}^{k} n_j^2 v_{1j}$ e a estatística de teste de Gehan é então $W = U_G^2 / V_G$. Sob a validade de H_0, W tem distribuição assintótica de Qui-quadrado com 1 grau de liberdade.

Note-se que cada diferença $(d_{ij} - e_{1j})$ é ponderada por n_j, o número de indivíduos em risco no instante t_j. Assim sendo, é atribuído maior peso às diferenças $(d_{1j} - e_{1j})$ correspondentes aos instantes onde o número total de indivíduos em risco é elevado, ou seja, aos instantes na parte inicial do estudo. Por isso, este teste é menos sensível que o teste long-rank a diferenças entre o número observado e o número esperado de acontecimentos ocorridos que se verifiquem na causa direita da distribuição do tempo de vida.

2.6 Modelo de regressão de Cox

Em 1972, Cox propôs um modelo que rapidamente se tornou no modelo de regressão mais utilizado na análise de tempos de vida, devido à sua flexibilidade e versatilidade (Cox, 1972).

Um dos aspetos inovadores do modelo de Cox reside no facto de ser formulado com base na relação entre a função de risco e as covariáveis. Com efeito, seja T uma variável aleatória contínua que representa o tempo de vida. A função de risco, no instante t e para um indivíduo a que esteja associado o vector de covariáveis $\mathbf{z} = (z_1, ..., z_p)'$, toma a forma:

$$\begin{aligned} h(t;\mathbf{z}) &= h_0(t)\exp(\beta'\mathbf{z}) \\ &= h_0(t)\exp(\beta_1 z_1 + ... + \beta_p z_p), \end{aligned} \tag{2.11}$$

em que $\beta_1, ..., \beta_p$ são os coeficientes de regressão (desconhecidos) que representam o efeito das covariáveis na sobrevivência e $h_0(t)$ é uma função arbitrária não negativa, que representa a função de risco para um indivíduo a que está associado o vetor $\boldsymbol{z} = 0$.

É um modelo de regressão semi-paramétrico visto que, embora o efeito das covariáveis seja modelado parametricamente, a função de risco subjacente $h_0(t)$, que descreve a forma comum das distribuições do tempo de vida para os indivíduos em estudo, não é especificada. Tal facto contribui para a flexibilidade do modelo.

Trata-se de um modelo de riscos proporcionais, visto que as funções de risco correspondentes a dois indivíduos com covariáveis $\boldsymbol{z_1}$ e $\boldsymbol{z_2}$ são proporcionais. De facto,

$$\frac{h(t;\boldsymbol{z_1})}{h(t;\boldsymbol{z_2})} = \exp\left\{\beta'(\boldsymbol{z_1} - \boldsymbol{z_2})\right\} \tag{2.12}$$

não depende de t.

Interpretação dos coeficientes

Habitualmente, esta interpretação não é feita em termos de β_j, mas sim de $\exp(\beta_j)$, por esta quantidade ter um significado mais direto no que diz respeito ao risco de morte.

Considere-se dois indivíduos a que estão associados os vectores de covariáveis $\boldsymbol{z_1}$ e $\boldsymbol{z_2}$, que diferem apenas nos valores da covariável z_j . Dada a forma da função de risco, tem-se então que

$$\begin{aligned} \frac{h(t;\mathbf{z}_1)}{h(t;\mathbf{z}_2)} &= \frac{h_0(t)\exp(\beta_1 z_{11} + \ldots + \beta_j z_{1j} + \ldots + \beta_p z_{1p})}{h_0(t)\exp(\beta_1 z_{21} + \ldots + \beta_j z_{2j} + \ldots + \beta_p z_{2p})} \\ &= \exp(\beta_j(z_{1j} - z_{2j})). \end{aligned} \tag{2.13}$$

Assim sendo, $\exp(\beta_j)$ representa o risco relativo de ocorrência do acontecimento para dois indivíduos que diferem de uma unidade nos valores da covariável z_j, sendo iguais os respetivos valores das restantes covariáveis.

Função de Verosimilhança

Suponha-se que se encontram em estudo n indivíduos e que foram observados k tempos de vida distintos $t_{(1)} < ... < t_{(k)}$, $k < n$. Seja

$$R_i = R(t_{(i)}) = \left\{ j : t_j \geq t_{(i)} \right\} \tag{2.14}$$

o conjunto de risco no instante $t_{(i)}$, isto é, o conjunto de índices associados aos indivíduos em observação imediatamente antes do instante $t_{(i)}$. Seja $z_{(i)}$ o vector de covariáveis associado ao indivíduo que morre em $t_{(i)}$.

Cox baseou a inferência sobre β na seguinte função utilizada como função de verosimilhança:

$$L(\beta) = \prod_{i=1}^{k} \frac{\exp(\beta' \mathbf{z}_{(i)})}{\sum_{l \in R_i} \exp(\beta' \mathbf{z}_l)}. \tag{2.15}$$

Esta função não depende de $h_0(t)$ e permite assim a realização de inferência sobre o vector de parâmetros β, sem que seja necessário fazer qualquer restrição à forma de $h_0(\cdot)$.

Note-se que a função (2.15) não é uma verosimilhança no sentido usual, pois não representa a probabilidade de realização de um acontecimento observável. Cox (1975) argumentou que (2.15) pode ser interpretada como uma verosimilhança parcial (destinada a permitir a realização de inferência na presença de parâmetros perturbadores), sendo aqui $h_0(t)$ entendida como uma função perturbadora.

Sob condições de regularidade bastante gerais, o estimador de máxima veresomilhança parcial de β é consistente, assintoticamente normal com valor médio β e matriz de covariância de $I(\beta)^{-1}$, onde

$$I_{jk}(\beta) = -E\left(\frac{\partial^2 \log L}{\partial \beta_j \partial \beta_k} \right). \tag{2.16}$$

Métodos de seleção de variáveis

Numa análise de regressão é habitual pretender identificar quais as covariáveis que têm influência significativa na sobrevivência dos indivíduos, de entre todas as que foram registadas, por exemplo, num estudo clínico. De facto, o modelo de regressão final deverá ser parcimonioso, embora possam ser também incluídas variáveis com relevância clínica que não se tenham revelado estatisticamente significativas, como variáveis biológicas de importância numa perspectiva clínica. Dado que o coeficiente β_j representa o efeito da covariável z_j na sobrevivência do indivíduo, para avaliar se existe evidência de que essa covariável influencia significativamente o tempo de vida, pode-se testar

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

utilizando o teste de Wald, em que a estatística de teste tem a forma de

$$Q = \frac{\hat{\beta}_j^2}{var(\hat{\beta}_j)} \sim \chi_1^2, \text{se } H_0 \text{ verdadeira.} \tag{2.17}$$

Note-se que se está a testar a hipótese de que a covariável z_j não tem influência significativa na sobrevivência, na presença das restantes covariáveis. Em geral, as estimativas $\hat{\beta}_j$ não são independentes umas das outras, o que dificulta a interpretação dos resultados de testes sobre os coeficientes associados a covaráveis incluídas num modelo. É preferível recorrer a métodos que permitam a comparação de modelos alternativos.

Considera-se um modelo de Cox com p covariáveis (modelo 1) e um modelo de Cox em que estão incluídas q covariáveis adicionais (modelo 2):

$$\text{Modelo 1: } h_0(t)\exp(\beta_1 z_1 + ... + \beta_p z_p).$$
$$\text{Modelo 2: } h_0(t)\exp(\beta_1 z_1 + ... + \beta_p z_p + \beta_{p+1} z_{p+1} + ... + \beta_{p+q} z_{p+q}).$$

O interesse é saber se os q termos adicionais incluídos no modelo 2 melhoram significativamente o poder explanatório deste modelo, relativamente ao modelo 1. Se tal não acontecer, os q termos podem ser omitidos e o modelo 1 é considerado adequado.

A função de verosimilhança resume a informação contida nos dados acerca dos parâmetros desconhecidos num dado modelo. Por isso, uma boa estatística que mede quão bem um modelo se ajusta aos dados é o valor da função de verosimilhança quando os parâmetros são substituídos pelas suas estimativas de máxima verosimilhança.

No entanto, a estatística $-2\log\hat{L}$ não pode ser usada por si só como medida da adequabilidade do modelo, pois o valor de $\hat{L}$ depende da dimensão da amostra. Logo, $-2\log\hat{L}$ só é útil ao comparar modelos ajustados aos mesmo dados.

Assim, os modelos podem ser comparados com base na diferença entre os valores da estatística $-2\log\hat{L}$ para cada modelo. Faz-se então um teste de razão de verosimilhanças para testar

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

Sob a hipótese nula

$$-2\log(\hat{L}_1/\hat{L}_2) \sim \chi_q^2. \tag{2.18}$$

O nível de significância considerado para a inclusão ou omissão de covariáveis não deve ser muito pequeno, recomenda-se que se utilize $\alpha \simeq 0.1$ (Collett, 2015).

2.7 Análise de resíduos

Para se poder avaliar a adequabilidade de um modelo de regressão, é fundamental uma definição apropriada de resíduo. A existência de observações censuradas e a própria forma do modelo de Cox levam a que não se possa fazer para este modelo uma definição análoga à de resíduo na regressão linear. De facto, a definição de resíduos é mais difícil e menos direta na modelação do tempo de vida do que no contexto de outros modelos de regressão.

Resíduos de Cox-Snell

Os primeiros resíduos propostos para o modelo de Cox foram os resíduos de Cox-Snell. Este tipo de resíduos são baseados na ideia de que, se o modelo é correto, os resíduos devem comportar-se como uma amostra proveniente de uma determinada distribuição conhecida.

Seja T uma variável aleatória contínua com função de distribuição F e função de sobrevivência S. Então $F(T) \sim U(0,1)$ e também $S(T) \sim U(0,1)$. Como $H(t) = -\log S(T)$, vem que $H(T)$ tem distribuição exponencial de valor médio 1.

Então, como no modelo de Cox se tem que

$$H(t;\boldsymbol{z}) = \int_0^t h_0(u)\exp(\boldsymbol{\beta}'\boldsymbol{z})du = \exp(\boldsymbol{\beta}'\boldsymbol{z})H_0(t), \tag{2.19}$$

o resíduo para o i-ésimo indivíduo, $i = 1,\ldots,n$, é definido como

$$r_i = \hat{H}(t_i) = \exp(\hat{\boldsymbol{\beta}}'\boldsymbol{z_i})\hat{H}_0(t_i), \tag{2.20}$$

em que $\hat{\beta}$ e $\hat{H}_0(t)$ são as estimativas de máxima verosimilhança parcial. Se o modelo que foi ajustado aos dados é satisfatório, então os valores estimados $\hat{H}(t_i)$ terão propriedades semelhantes aos verdadeiros valores de $H(t_i)$. Portanto, os resíduos r_i devem comportar-se, aproximadamente, como uma amostra aleatória proveniente de uma população com distribuição Exp(1).

Quando existem dados censurados é necessário ter isso em conta, visto que se determinada observação é censurada também é o resíduo correspondente.

Os resíduos de Cox-Snell têm propriedades bastante diferentes dos resíduos usados em regressão linear porque, por exemplo, não se distribuem de forma simétrica em torno de zero. Para além disso, como os resíduos de Cox-Snell seguem uma distribuição exponencial quando o modelo que foi ajustado aos dados é apropriado, terão uma distribuição bastante assimétrica.

Uma desvantagem destes resíduos é que não indicam qual o tipo de afastamento do modelo. Devem ser usados com cautela, principalmente no caso de pequenas amostras, visto que o afastamento da

distribuição dos resíduos da distribuição exponencial pode ser devido, em parte, à substituição de β e $H_0(t)$ pelas suas estimativas de máxima verosimilhança parcial.

Resíduos de Schoenfeld

Os resíduos de Schoenfeld definem resíduos que não dependem do tempo para que o i-ésimo resíduo possa ser representado graficamente contra t_i para avaliar o pressuposto de riscos proporcionais (Schoenfeld, 1982).

Estes resíduos foram propostos por Schoenfeld (1982) e diferem dos resíduos de Cox-Snell no sentido de não ser necessário obter uma estimativa da função de risco cumulativa e que a cada indivíduo corresponde um conjunto de valores, um por cada covariável que foi incluída no modelo de regressão de Cox.

Para o i-ésimo indivíduo em estudo, o resíduo de Schoenfeld correspondente à covariável $z_j, j = 1, ..., p$ é dado por

$$r_{ji} = \delta_i(z_{ji} - a_{ji}) \tag{2.21}$$

onde

$$\delta_i = \begin{cases} 1 & \text{se } t_i \text{ é uma observação não censurada} \\ 0 & \text{se } t_i \text{ é uma observação censurada} \end{cases} \tag{2.22}$$

e

$$a_{ji} = \frac{\sum_{l \in R_i} z_{jl} \exp(\hat{\beta}' \mathbf{z}_l)}{\sum_{l \in R_i} \exp(\hat{\beta}' \mathbf{z}_l)}. \tag{2.23}$$

Para um indivíduo cujo tempo de vida não foi observado, estes resíduos são sempre nulos. Então, para distinguir resíduos que são verdadeiramente iguais a zero daqueles que correspondem a observações censuradas, estes últimos são geralmente indicados como valores omissos.

Para um indivíduo cuja morte foi observada em t_i, o resíduo é a diferença entre o valor da covariável z_j correspondente a esse indivíduo e uma média ponderada dos valores dessa variável para todos os indivíduos em risco em t_i. O peso associado a um indivíduo $l \varepsilon R_i$ é $\exp(\hat{\beta}' \mathbf{z_l})$. Note-se que

$$\frac{\partial \log L}{\partial \beta_j} = \sum_{i=1}^{n} \delta_i \left\{ z_{ji} - \frac{\sum_{l \in R_i} z_{jl} \exp(\beta' \mathbf{z}_l)}{\sum_{l \in R_i} \exp(\beta' \mathbf{z}_l)} \right\}, \tag{2.24}$$

onde L é a função de verosimilhança parcial. A i-ésima parcela desta soma, calculada em $\hat{\boldsymbol{\beta}}$, é então o resíduo de Schoenfeld correspondente à covariável z_j para o i-ésimo indivíduo.

Como as estimativas $\hat{\beta}_j$ são tais que

$$\left.\frac{\partial \log L}{\partial \beta_j}\right|_{\hat{\beta}} = 0, \tag{2.25}$$

conclui-se que a soma, para todos os indivíduos em estudo, dos resíduos de Schoenfeld correspondentes a cada covariável, é igual a zero.

Resíduos Martingala

Estes resíduos mostram-se extremamente úteis para a determinação da forma funcional que deve ser usada para uma dada covariável, de modo a explicar o melhor possível no seu efeito na sobrevivência, bem como na deteção de *outliers*.

Quando todas as covariáveis são fixas, no início do estudo, o resíduo martingala associado ao i-ésimo indivíduo, $i = 1, ..., n$, é dado por

$$\hat{M}_i = \delta_i - \exp(\hat{\boldsymbol{\beta}}' \boldsymbol{z_i}) \hat{H}_0(t_i) = \delta_i - r_i$$

em que δ_i é a variável indicatriz usual e r_i é o resíduo de Cox Snell associado a esse indivíduo.

Os resíduos martingala apresentam grande assimetria e tomam valores no intervalo $(-\infty, 1)$, sendo negativos os resíduos correspondentes a observações censuradas.

2.8 Riscos Competitivos

Os métodos anteriormente referidos, que podem ser considerados métodos clássicos da análise de sobrevivência, como o estimador de Kaplan-Meier e o modelo de regressão de Cox, têm o pressuposto que a censura existente é independente ou não informativa. Isto significa que os indivíduos a que correspondem observações censuradas têm o mesmo risco de sofrer o acontecimento de interesse no futuro como aqueles que ainda se encontram no estudo (Austin et al., 2016; Putter et al., 2007). Estes métodos consideram apenas um único acontecimento de interesse. No entanto, existem situações nas quais vários acontecimentos se podem considerar igualmente relevantes para o problema em estudo (Rocha & Papoila, 2009). A teoria de riscos competitivos é adequada nestas situações em que cada indivíduo está sujeito a dois ou mais acontecimentos de interesse.

É de salientar que Daniel Bernoulli publicou em 1766 um ensaio motivado pela controvérsia sobre a vacina da varíola, onde são introduzidas as bases da teoria dos riscos competitivos, como pode ser consultado em Klein et al., 2016.

As situações em que existem riscos competitivos, são expressas de formas diferentes consoante diferentes autores. Kalbfleich e Prentice (2002) descreveram os riscos competitivos como situações

em que um indivíduo pode sofrer mais de que um acontecimento de interesse. Geskus (2015) define o conceito de riscos competitivos como a situação onde um acontecimento impede a ocorrência de outro acontecimento ou altera drasticamente a probabilidade de ocorrência deste outro acontecimento.

Em estudos epidemiológicos e clínicos, os riscos competitivos são frequentemente ignorados na análise de dados de sobrevivência. No entanto, a não consideração de riscos competitivos, geralmente, leva a uma sobrestimação da incidência cumulativa do acontecimento de interesse (Lau et al., 2009; Putter et al., 2007; Satagopan et al., 2004).

Neste trabalho, os acontecimentos de interesse são morte por COVID-19, morte por outra causa e recuperação. Quando um indivíduo sofre o primeiro destes acontecimentos, termina o seu tempo de observação. Salienta-se que quando um indivíduo morre por COVID-19, já não pode recuperar ou morrer por outra causa, sendo portanto nula a probabilidade de ocorrência dos restantes acontecimentos. O mesmo acontece quando o indivíduo morre por outra causa. Quanto à recuperação, durante o período do estudo, não existiu qualquer tipo de reincidência de infeção pelo SARS-CoV-2, por isso, se o indivíduo recupera da COVID-19, já não pode morrer por essa causa e como o seu tempo de observação termina quando se dá a recuperação, não se sabe se o indivíduo em causa morreu por outra causa.

2.8.1 Funções específicas da causa

Existem várias funções importantes na descrição de riscos competitivos, como por exemplo, a subdistribuição, subdensidade, subrisco, risco específica da causa e risco da subdistribuição. Porém, existe discordância em termos de nomenclatura, na literatura, dificultando a pesquisa e a leitura (Pintilie, 2006). Por exemplo, a função de subdistribuição tem sido referida como função de incidência cumulativa (Kalbfleisch & Prentice, 2002), distribuição marginal (Pepe & Mori, 1993), risco absoluto específico da causa (Benichou & Gail, 1990) e probabilidade de falha específica da causa (Gaynor et al., 1993). Neste trabalho, os termos função de incidência cumulativa (CIF) e função de risco específica da causa serão utilizados.

Suponha-se que os indivíduos de uma população estão sujeitos a m causas de morte ou a m acontecimentos de interesse. Quando ocorre uma morte, observa-se o tempo de vida T e a causa de morte J, $J \in \{1,2,...,m\}$. Esta abordagem descreve o problema em termos das funções de risco específicas da causa (Kalbfleisch & Prentice, 2002).

A função de risco específica da causa ou função de subrisco para o j-ésimo acontecimento, $\lambda_j(t)$, para $j = 1,2,...,m$, é definida por

$$\lambda_j(t) = \lim_{dt \to 0} \left\{ \frac{P(t < T \leq t + dt, J = j | T > t)}{dt} \right\} \tag{2.26}$$

e descreve a probabilidade instantânea de morte devida à causa j no instante t, na presença das outras causas de morte.

Pela definição de $\lambda_j(t)$ em (2.26) e recorrendo ao método usado para deduzir (2.3), tem-se

$$\begin{aligned}\lambda_j(t) &= \lim_{dt\to 0}\left\{\frac{P(t<T\leq t+dt, J=j)}{dtP(T>t)}\right\} \\ &= \frac{1}{P(T>t)}\lim_{dt\to 0}\left\{\frac{P(t<T\leq t+dt, J=j)}{dt}\right\} \\ &= \frac{f_j}{S(t)},\end{aligned} \tag{2.27}$$

Como

$$P(t<T\leq t+dt|T>t) = \textstyle\sum_{j=1}^{m} P(t<T\leq t+dt, J=j|T>t),$$

tem-se que a função de risco global é dada por

$$\lambda(t) = \lim_{dt\to 0}\left\{\frac{P(t<T<t+dt|T>t)}{dt}\right\} \tag{2.28}$$

e satisfaz a relação

$$\lambda(t) = \sum_{j=1}^{m}\lambda_j(t).$$

Do mesmo modo, a função de risco cumulativa global é

$$\Lambda(t) = \sum_{j=1}^{m}\Lambda_j(t)$$

onde $\Lambda_j(t)$ é a função de risco cumulativa para a j-ésima causa é dada por

$$\Lambda_j(t) = \int_0^t \lambda_j(x)dx = \int_0^t \left\{f_j(x)/S(x)\right\}dx.$$

Logo, a função de sobrevivência do tempo de vida T pode ser representada por

$$S(t) = \exp\left(-\sum_{j=1}^{m}\int_0^t \lambda_j(u)du\right).$$

A função de sobrevivência específica da causa j ($j = 1,2,...,m$) é dada por

$$S_j(t) = P(T>t, J=j) = \int_t^{\infty}\lambda_j(u)S(u)du.$$

Note-se que

$$\begin{aligned} S(t) &= P(T > t) \\ &= \sum_{j=1}^{m} P(T > t, J = j) \\ &= \sum_{j=1}^{m} S_j(t). \end{aligned}$$

Apesar de $S(t)$ também poder ser escrita como

$$S(t) = \prod_{j=1}^{m} S_j^*(t),$$

onde

$$S_j^*(t) = \exp\left(-\int_0^t \lambda_j(u)du\right) \qquad j = 1, ..., m,$$

as funções $S_j^*(t)$ não são observáveis. Isto deve-se ao facto de nunca se saber qual é o acontecimento que pode ocorrer depois do instante t, quando existe mais de um acontecimento possível.

Note-se que estas funções não podem, geralmente, ser interpretadas como funções de sobrevivência para $m > 1$.

Pode-se ainda definir a função de incidência cumulativa da causa j como sendo

$$F_j(t) = P(T \leq t, J = j) = \int_0^t \lambda_j(u) S(u) du, \tag{2.29}$$

para $j = 1, 2, ..., m$. O valor máximo que a função da incidência cumulativa pode atingir é

$$P(T < \infty, J = j) = P(J = j) = \pi_j$$

onde π_j é a probabilidade do acontecimento j ocorrer. Consequentemente, $F_j(t) \to \pi_j$ se $t \to \infty$ e visto que $F_j(t)$ não tende para zero, $F_j(t)$ não é uma função de distribuição própria. Tendo esta afirmação em conta, a função de incidência cumulativa é também referida como função de subdistribuição, como referido anteriormente.

A função de distribuição global, $F(t)$, corresponde à probabilidade de ocorrência de um qualquer acontecimento até ao instante t e é igual à soma das funções de incidência cumulativa para cada acontecimento, i.e.,

$$\begin{aligned} F(t) &= P(T \leq t) \\ &= \sum_{j=1}^{m} P(T \leq t, J = j) \\ &= \sum_{j=1}^{m} F_j(t). \end{aligned}$$

Note-se também que

$$F_j(t) + S_j(t) = P(J = j).$$

A função de densidade para o j-ésimo acontecimento é definida como

$$f_j(t) = \frac{\partial F_j(t)}{\partial t}.$$

A função de risco da subdistribuição (ou função de risco associada à função de incidência cumulativa) da causa j, para $j = 1, 2, ..., m$ (Fine & Gray, 1999) , pode ser definida como

$$\gamma_j(t) = \lim_{dt \to 0} \left\{ \frac{P(t < T < t + dt, J = j | T > t \text{ ou } (T \leq t \text{ e } J \neq j))}{dt} \right\}. \tag{2.30}$$

A relação entre $\gamma_j(t)$ e as funções de subdensidade e de subdistribuição pode ser expressa como

$$\gamma_j(t) = \frac{f_j(t)}{1 - F_j(t)}.$$

2.8.2 Estimador de função de incidência cumulativa

A utilização do estimador de Kaplan-Meier, considerando como censuradas as observações correspondentes aos indivíduos a quem ocorreu outro que não o "acontecimento de interesse", pode levar a uma estimativa da incidência cumulativa que sobrestima substancialmente a incidência do referido acontecimento, em particular, quando o acontecimento competitivo ocorre com maior frequência do que o acontecimento de interesse (Wolbers et al., 2014). Esta sobrestimação será explicada no subsecção seguinte.

Assim sendo, o estimador de Kaplan-Meier foi generalizado de forma a ser utilizado num problema de riscos competitivos. Considera-se então os instantes de morte distintos $t_{j1} < t_{j2} < ... < t_{jkj}$ onde k_j é o número de instantes de morte devida à causa j para $j = 1, ..., m$. Suponha-se ainda que o acontecimento

de tipo j ou morte devida à causa j ocorre d_{ji} vezes no instante t_{ji}, $i = 1, ..., k_j$. O estimador de máxima verosimilhança não paramétrico de S_j é

$$\hat{S}_j(t) = \prod_{i:t_{ji} \leq t} \left(1 - \frac{d_{ji}}{n_{ji}}\right), \tag{2.31}$$

em que n_{ji} é o número de indivíduos em risco no instante t_{ji}. Ignorando as causas de morte, o estimador de Kaplan-Meier da função de sobrevivência global é $\hat{S}_j(t) = \prod_m^{j=1} \hat{S}_j(t)$, desde que não haja observações empatadas entre instantes de morte de tipos diferentes. O estimador da função de incidência cumulativa da causa j para $j = 1, ..., m$ é dado por

$$\hat{F}_j(t) = \sum_{i:t_i \leq t} \frac{d_{ji}}{n_i} \hat{S}(t_{i-1}). \tag{2.32}$$

Obter a variância exata do estimador da função de incidência cumulativa da causa j, para $j = 1, 2, ..., m$, é uma tarefa difícil, pois é preciso avaliar

$$\begin{aligned} \text{Var}\left(\hat{F}_j(t)\right) = &\sum_{t_i \leq t} \text{Var}\left(\left\{d_{ji}/n_i\right\} \hat{S}(t_i)\right) \\ &+ 2 \sum_{t_i < t} \sum_{t_v > t_i} \text{Cov}\left(\left\{d_{ji}/n_j\right\} \hat{S}(t_i), \left\{d_{jv}/n_v\right\} \hat{S}(t_v)\right). \end{aligned}$$

Recorrendo ao método de Aalen, $\text{Var}\left(\hat{F}_j(t)\right)$ pode ser estimada como:

$$\begin{aligned} \widehat{\text{Var}}\left(\hat{F}_j(t)\right) = &\sum_{t_i \leq t} \left\{ \left[\hat{F}_j(t) - \hat{F}_j(t_j)\right]^2 \frac{d_i}{(n_i - 1)(n_i - d_i)} \right\} \\ &+ \sum_{t_i \leq t} \hat{S}(t_{i-1})^2 \frac{d_{ji}(n_i - d_{ji})}{n_i^2 (n_i - 1)} \\ &- 2 \sum_{t_i \leq t} \left[\hat{F}_j(t) - \hat{F}_j(t_i)\right] \hat{S}(t_{i-1}) \frac{d_{ji}(n_i - d_{ji})}{n_i (n_i - d_i)(n_i - 1)}. \end{aligned} \tag{2.33}$$

Visto que $\hat{F}_j$ é uma função em escada, os termos que incluem $\hat{F}_j(t) - \hat{F}_j(t_i)$ são zero, exceto se um qualquer acontecimento j ocorreu entre os instantes t_i e t. Da mesma forma, os termos que envolvam d_{ji} serão zero com exceção nos instantes em que o acontecimento j foi registado. Portanto, o último termo de cada soma é diferente de zero se, no instante t, o acontecimento j tiver ocorrido. Quando existe apenas um acontecimento de interesse e na ausência de riscos competitivos, a expressão (2.33) para a estimativa da variância é semelhante à fórmula de Greenwood representada em (2.8).

2.8.3 Estimador de Kaplan-Meier e riscos competitivos

Na análise de sobrevivência, quando há um único acontecimento de interesse e na ausência de riscos competitivos, o estimador de Kaplan-Meier é frequentemente usado na descrição do tempo

até à ocorrência do acontecimento dos indivíduos em estudo. É simples de calcular, de representar graficamente e é facilmente interpretado por investigadores clínicos e outros. Parece natural aplicar o mesmo método na presença de riscos competitivos. Na realidade, muitos trabalhos de investigação aplicam a estimativa de Kaplan-Meier neste tipo de situação. No entanto, a sua interpretação na presença de riscos competitivos é diferente (Pintilie, 2006).

Este estimador é baseado na premissa que a censura é não informativa, isto é, censura independente do acontecimento de interesse. Porém, na presença de riscos competitivos existe censura informativa.

Quando se aplica o estimador Kaplan-Meier a riscos competitivos, os restantes acontecimentos de interesse são ignorados. Portanto, o estimador de Kaplan-Meier estima para o j-ésimo acontecimento, a probabilidade desse acontecimento ocorrer depois do instante t como se o acontecimento j fosse o único acontecimento de interesse, ou seja, numa situação irrealista em que os outros riscos foram eliminados.

Para além disso, o estimador de Kaplan-Meier assume que o indivíduo irá inevitavelmente sofrer de um determinado acontecimento, não havendo a possibilidade de que o indivíduo pode nunca sofrer do acontecimento em causa.

Pode-se provar que (Pintilie, 2006), para qualquer instante t_i, o complemento da estimativa de Kaplan-Meier $(1-KM)$ é maior do que a estimativa da função de incidência cumulativa. Sem qualquer perda de generalidade, suponha-se que há dois tipos de acontecimento: o acontecimento de interesse, representado pelo índice 1 e o acontecimento competitivo representado pelo índice 2. Algebricamente, a estimativa de Kaplan-Meier pode ser escrita como

$$\begin{aligned} KM_1(t_i) &= \prod_{j=1}^{i} \frac{n_j - d_{1j}}{n_j} \\ &= \frac{n_i - d_{1i}}{n_i} KM_1(t_{i-1}) \\ &= KM_1(t_{i-1}) - \frac{d_{1i}}{n_i} KM_1(t_{i-1}). \end{aligned}$$

Isto implica que

$$\begin{aligned} 1 - KM_1(t_i) &= 1 - KM_1(t_{i-1}) + \frac{d_{1i}}{n_i} KM_1(t_{i-1}) \\ &= \sum_{j=1}^{i} \frac{d_{1j}}{n_j} KM_1(t_{j-1}) \end{aligned}$$

Esta expressão pode ser comparada com (2.32). O valor de KM_1 é a estimativa de Kaplan-Meier quando existe apenas um acontecimento de interesse, enquanto $\hat{S}$ é a estimativa de Kaplan-Meier quando o acontecimento de interesse e os riscos competitivos são considerados. Assim, $\hat{S}(t) \leq KM_1(t)$ para qualquer t. A igualdade acontece quando não existem riscos competitivos. Concluí-se que $\hat{F}_1(t_i) \leq 1 - KM_1(t_i)$.

2.8.4 Testes para comparação de grupos

Teste de Gray

Sem perda da generalidade, apenas dois acontecimentos serão considerados: o acontecimento de interesse e o acontecimento competitivo. O teste de Gray permite comparar as funções de risco associadas às funções de incidência cumulativa de k populações diferentes (Gray, 1988). Tendo em conta que as funções envolvidas no teste referem-se ao acontecimento de interesse, o índice para o tipo de acontecimento será suprimido.

A hipótese nula em questão é

$$H_0 : F_i(t) = F_j(t), \forall_{i,j} = 1, ..., k, i \neq j,$$

onde $F_i(t)$ é a função de incidência cumulativa relativa ao acontecimento de interesse na população i. O teste para k-amostras introduzido por Gray em 1988, compara as médias ponderadas das funções de risco da subdistribuição para o acontecimento de interesse. A forma geral da pontuação para o grupo i é

$$z_i = \int_0^\tau W_i(t) \{\gamma_i(t) - \gamma_0(t)\} dt, \tag{2.34}$$

onde τ é o tempo máximo observado nos dois grupos, $W_i(t)$ é uma função peso, $\gamma_i(t)$ é a função de subdistribuição para o grupo i e $\gamma_0(t)$ é a função de subdistribuição para os grupos em conjunto. Normalmente a função peso está na forma $W_i(t) = L(t)R_i(t)$ onde $L(t)$ é uma função do tempo e

$$R_i(t) = n_i(t) \frac{1 - \hat{F}_i(t-)}{\hat{S}_i(t-)}, \tag{2.35}$$

onde $n_i(t)$ é o número de indivíduos em risco no instante t no grupo i, $F(t-)$ corresponde ao limite à esquerda da função de incidência cumulativa do acontecimento de interesse e $S(t-)$ é o limite à esquerda da probabilidade de não sofrer nenhum acontecimento, que é estimado com o método de Kaplan-Meier. Desta forma, R_i representa um número ajustado de indivíduos em risco.

A estatística de teste baseada em (2.34) segue, sob validade da hipótese nula, uma distribuição χ^2_{k-1}.

Teste de Pepe e Mori

Este teste permite comparar, de uma forma direta, as funções de incidência cumulativa de duas populações diferentes (Pepe, 1991). Pode-se mostrar que, sob a validade da hipótese nula, $H_0 : F_i(t) = F_j(t), \forall_{i,j} = 1, ..., k, i \neq j$, a estatística de teste

$$S = \sqrt{\frac{N_1 N_2}{N_1 + N_2}} \int_0^{\tau} W(t) \{F_1(t) - F_2(t)\} dt \tag{2.36}$$

tem distribuição assintótica Normal de valor médio 0 e variância σ^2. Em (2.36) note-se que N_i é o número total de indivíduos no grupo i, $F_j(t)$ é a função de incidência cumulativa para a população j e $W(t)$ é uma função peso que tem a seguinte expressão:

$$W(t_j) = \frac{(N_1 + N_2)\,\hat{G}_1(t_{j-1})\,\hat{G}_2(t_{j-1})}{N_1 \hat{G}_1(t_{j-1}) + N_2 \hat{G}_2(t_{j-1})},$$

onde $\hat{G}_i(t)$ é o estimador de Kaplan-Meier da função de sobrevivência do tempo de censura da população i, considerando que as mortes correspondem às observações censuradas ou aos acontecimentos competitivos. A função peso é decrescente no que diz respeito ao tempo, por isso, o peso associado à diferença entre as duas funções de incidência cumulativa diminui à medida que o tempo aumenta.

A variância de S é estimada pela média ponderada das estimativas das variâncias nos dois grupos, isto é,

$$\hat{\sigma}^2 = \frac{N_1 N_2 \left(\hat{\sigma_1}^2 + \hat{\sigma_2}^2\right)}{N_1 + N_2},$$

com

$$\hat{\sigma_i}^2 = \sum_j \frac{\left[\nu_1(t_j) - \hat{F_{rc}}(t_j)\nu_2(t_j)\right]^2 d_{evj} + \nu_2^2(t_j)(d_j - d_{evj})}{n_j(n_j - 1)},$$

onde n_j é o número de indivíduos em risco no instante t_j no grupo i, d_j é o número de acontecimentos (de interesse ou riscos competitivos) no instante t_j no grupo i e d_{evj} refere-se ao número de acontecimentos de interesse no tempo t_j no grupo i. Da mesma forma, F é a função de incidência cumulativa para o acontecimento de interesse para o grupo i, F_{rc} é a função de incidência cumulativa para o risco competitivo para o grupo i e σ é o desvio padrão para o grupo i.

Note-se que

$$\nu_1(t_j) = \sum_{t_k \geq t_j} W(t_k)(t_{k+1} - t_k)\left(1 - \hat{F}_i(t_k)\right)$$
$$\nu_2(t_j) = \sum_{t_k \geq t_j} W(t_k)(t_{k+1} - t_k).$$

2.8.5 Modelo de regressão de Fine e Gray

Depois de averiguar se existem diferenças significativas entre as funções de incidência cumulativa, para cada uma das variáveis explicativas isoladamente, pelos métodos anteriormente referidos (Teste de Gray e Teste de Pepe e Mori), seria importante analisar quais dessas variáveis têm influência nos tempos até à ocorrência dos acontecimentos. Assim, Fine e Gray (1999) propuseram um modelo de regressão de riscos proporcionais, modelando o efeito das variáveis explicativas através da função de risco associada à função incidência cumulativa, permitindo identificar os fatores de prognóstico que estão associados a cada um dos acontecimentos.

Suponha-se que para cada indivíduo está associado um vetor de variáveis explicativas $\boldsymbol{z} = (z_1, ..., z_p)'$. Fine e Gray (1999) propuseram um modelo de regressão que é uma generalização do modelo de Cox, de modo a permitir a sua aplicação na presença de riscos competitivos. Assim, no instante t, a função de risco associada à função de incidência cumulativa, dado $\boldsymbol{z}$ é definida por

$$\gamma(t;\boldsymbol{z}) = \gamma_0(t)\exp(\boldsymbol{\beta}'\boldsymbol{z}), \tag{2.37}$$

sendo $\gamma_0(t)$ a função de risco subjacente à função de incidência cumulativa e $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)'$ o vetor dos p coeficientes de regressão.

Assume-se que os k instantes distintos de morte onde se observou o acontecimento de interesse são $t_1 < t_2 < ... < t_k$. A função de verosimilhança parcial para a estimação dos coeficientes $\boldsymbol{\beta}$, é definida por

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{k} \frac{\exp(\boldsymbol{\beta}'\boldsymbol{z}_i)}{\sum_{j\in R_i} W_{ij}\exp(\boldsymbol{\beta}'\boldsymbol{z}_j)}$$

em que W_{ij} é um peso e R_i designa o conjuntos de indivíduos em risco em t_i, que é constituído pelos indivíduos aos quais não ocorreu nenhum dos acontecimentos possíveis, bem como por aqueles a quem aconteceu o acontecimento competitivo, até ao instante t_i. Desta forma, o conjuntos de risco no instante t_i, é dado por $R_i = \{\, j : T_j > t_i$ ou ($T_j < t_i$ e ter ocorrido ao indivíduo o acontecimento competitivo) $\}$.

O peso W_{ij} é definido por

$$W_{ij} = \frac{\hat{G}(t_i)}{\hat{G}(min(t_i, t_j))}$$

onde $\hat{G}$ é o estimador de Kaplan-Meier da função de sobrevivência dos tempos de censura.

Validação do modelo

A adequabilidade do modelo específico da causa pode ser avaliada através dos métodos já referidos anteriormente, nomeadamente através da análise de resíduos.

Capítulo 3

Análise dos Dados

Este estudo analisa a informação sobre 360 914 indivíduos que foram registados no BI SINAVE por testarem positivo à COVID-19, em Portugal. Destes indivíduos, sabe-se se morreram por COVID-19, se morreram por outra causa ou se recuperaram. Os indivíduos que não sofreram nenhum dos acontecimentos referidos desde a sua data de entrada para o estudo até à data final, foram considerados casos ativos, ou seja, deram origem a observações censuradas à direita. O período abrangido por este estudo é desde 2 de março até 31 de dezembro de 2020.

A variável resposta representa o tempo desde a data em que o teste realizado foi positivo para SARS-CoV-2 até a ocorrência do primeiro acontecimento (morte por COVID-19, morte por outra causa ou recuperação). A data de fim de observação corresponde à data do óbito, ou à data em que se deu a recuperação ou à data final do estudo. O tempo em observação foi registado em dias.

É de notar que os 360 914 indivíduos não correspondem à totalidade de casos positivos à COVID-19 no período referido. De facto, durante um certo período de tempo não existia a obrigatoriedade de registo no SINAVE, de testes TRAg positivos realizados por estabelecimentos prestadores de cuidados de saúde, com registo válido na Entidade Reguladora da Saúde (ERS). Os autotestes também não eram sujeitos a obrigatoriedade de registo. Como referido anteriormente, também não foram contabilizados os casos em que os indivíduos não tinham indicação de sexo, idade e região.

O tratamento da base de dados proveniente da DGS e os métodos de análise estatística como construção de gráficos, aplicação de testes estatísticos e modelação foram feitos através do *software* R versão 4.1.2, recorrendo aos *packages* `survival`, `survminer`, `cmprsk` e `riskRegression`.

3.1 Análise exploratória

Tabela 3.1: Análise descritiva.

Variáveis	Amostra (N=360 914) (100%)	Morte por COVID-19 (N=5 197) (1,44%)	Morte por outra causa (N=411) (0,11%)	Recuperação (N=313 377) (86,83%)
Sexo				
Feminino	199 931 (55,34%)	2 466 (47,45%)	209 (50,85%)	174 129 (55,57%)
Masculino	160 983 (44,60%)	2 731 (52,55%)	202 (49,15%)	139 248 (44,43%)
Idade (mediana)	43	84	84	43
Grupo etário				
[0;20]	57 892 (16,04%)	3 (0,06%)	0	51 075 (16,30%)
]20;60]	219 691 (60,84%)	199 (3,83%)	31 (7,54%)	194 373 (62,03%)
]60;70]	34 031 (9,43%)	444 (8,54%)	49 (11,92%)	29 427 (9,39%)
]70;80]	22 462 (6,22%)	1165 (22,42%)	82 (19,95%)	18 416 (5,88%)
80+	26 838 (7,44%)	3 386 (65,15%)	249 (60,58%)	20 086 (6,41%)
ARS				
Norte	190 341 (52,74%)	2 479 (47,70%)	200 (48,66%)	170 267 (54,33%)
Centro	43 890 (12,16%)	799 (15,7%)	56 (13,63%)	35 494 (11,33%)
LVT	109 315 (30,29%)	1 695 (32,61%)	146 (35,52%)	94 052 (30,01%)
Alentejo	8 703 (2,41%)	149 (2,88%)	4 (0,97%)	6 518 (2,08%)
Algarve	6 385 (1,77%)	56 (1,08%)	3 (0,73%)	5 172 (1,65%)
Madeira	577 (0,16%)	2 (0,038%)	1 (0,24%)	464 (0,15%)
Açores	1703 (0,47%)	17 (0,33%)	1 (0,24%)	1 410 (0,45%)

Nesta amostra de 360 914 pessoas que testaram positivo à COVID-19, apenas 1,44% faleceram devido à doença, 0,11% faleceram devido a outra causa e 86,83% recuperaram, sendo que 11,62% correspondem aos casos ativos, isto é, a indivíduos que não sofreram nenhum acontecimento até ao final do estudo, ou seja, deram origem a observações censuradas à direita.

A mortalidade por COVID-19, em Portugal e no ano de 2020, não se mostrava muito preocupante até meados de novembro, altura em que o número diário de óbitos começou a aumentar, atingindo um pico de 103 óbitos no dia 12 de dezembro. A 30 de janeiro de 2021 foi registado o número mais elevado de mortes diárias por COVID-19 em Portugal, correspondendo a 297 óbitos (fonte: https://covid19.min-saude.pt/numero-de-novos-casos-e-obitos-por-dia). Neste estudo apenas constam registos até dia 31 de dezembro de 2020 e, por isso, não será abordado a fase mais preocupante da pandemia em Portugal.

Quanto às mortes por outras causas, a sua percentagem mostra-se inferior comparativamente à percentagem de mortes por COVID-19. Diz-se que um indivíduo com COVID-19 morre por outra causa quando a causa de morte não está claramente associada à doença COVID-19, por exemplo, se for de traumatismo ou neoplasia.

Sendo que a COVID-19 não apresenta uma elevada letalidade, esperava-se que a maioria dos indivíduos recuperasse. Os dados relativos aos recuperados são de baixa qualidade pois têm por base o grupo etário em que o indivíduo se encontra e em que vaga se deu a infeção pelo vírus SARS-CoV-2. São dados com pouco rigor, mas para uma análise de sobrevivência com riscos competitivos foi imperativo usar os dados disponíveis quanto à recuperação.

Esta amostra de doentes infetados pelo SARS-CoV-2, mostra-se equilibrada quanto ao sexo, isto é, a percentagem de indivíduos do sexo masculino é bastante próxima da percentagem dos indivíduos do sexo feminino, sendo $44,60\%$ e $55,34\%$ as respectivas percentagens.

Relativamente aos grupos etários, o grupo etário que apresenta uma frequência relativa mais elevada é o grupo que contém os indivíduos com idades compreendidas entre os 20 e os 60 anos e isso deve-se, em parte, por ser a classe etária com maior amplitude, mas também devido à distribuição da população pelos grupos etários em Portugal.

O grupo etário correspondente à população estudantil, mostrou-se mais resistente à evolução da doença grave, sendo pouco provável a ocorrência de morte nestas idades. Dos 20 aos 60 anos trata-se dos indivíduos pertencentes à população ativa, com uma percentagem baixa relativamente à ocorrência de morte, embora mais elevada que no grupo [0;20]. Nos restantes grupos etários nota-se um aumento de mortalidade. Sabe-se que com o envelhecimento progressivo da população, existe um aumento da prevalência de pessoas com doenças crónicas incapacitantes. Apesar de o envelhecimento não ser sinónimo de doença e dependência, existe uma considerável prevalência de doenças crónicas nos idosos em Portugal, sendo frequente a multimorbilidade (Portugal é o 3.º País da OCDE com maior percentagem de pessoas com mais de 65 anos a viver com duas ou mais doenças crónicas, segundo o relatório Health at a Glance 2019).

Todos os grupos etários mostraram um aumento de incidência a 14 dias por 100 mil habitantes a partir de outubro, devido à reabertura das escolas, fim de férias e aumento de testagem, como se pode observar na Figura 3.1 proveniente do Relatório Técnico da Vigilância da Infeção por SARS-CoV-2/ COVID-19.

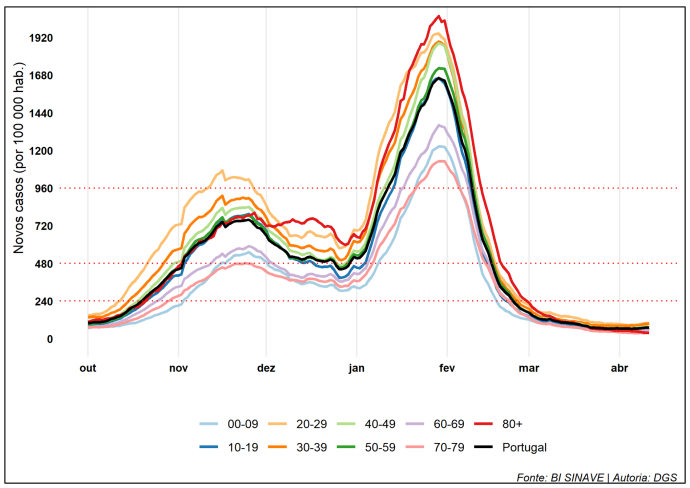


Figura 3.1: Incidência cumulativa a 14 dias (por 100 000 habitantes), por grupo etário, entre 01/10/2020 e 11/04/2021. Fonte: BI SINAVE.

Ao analisar a distribuição dos indivíduos pelas Administrações Regionais de Saúde, a ARS Norte e a ARS LVT apresentam os valores mais elevados sendo também as regiões mais populosas.

Tabela 3.2: Tempo, em dias, até ocorrência de um acontecimento.

Tempo até ao acontecimento	Mínimo	Máximo	Média	Mediana
Morte por COVID-19	1	268	11,15	9
Morte por outra causa	1	189	13,53	9
Recuperação	9	43	15,96	14

Ao considerar o tempo em observação, para os finais fatais, o tempo mínimo de observação é de um dia, correspondendo a muitas situações em que o indivíduo faleceu e só se soube o resultado do teste posteriormente à data do óbito. Quanto ao tempo máximo de observação, entre o dia de positividade à COVID-19 e a ocorrência de morte por COVID-19 ou morte por outra causa é de 268 e 189 dias, respectivamente.

Os critérios de recuperação à COVID-19 referidos na norma 004/2020 sofreram várias atualizações. Até 14 outubro de 2020, o tempo mínimo de recuperação era de 9 dias e o indivíduo só poderia ser considerado recuperado após a realização de dois testes à COVID-19 com resultado negativo. A partir de 14 outubro de 2020, inclusive, os indivíduos assintomáticos e sintomáticos ligeiros teriam de realizar um isolamento de 10 dias, sendo que nos últimos três dias consecutivos não poderiam ter sintomas e os indivíduos sintomáticos graves estariam em isolamento 20 dias, sendo que nos últimos três dias consecutivos não poderiam ter sintomas.

3.2 Inferência estatística não paramétrica

Estimativa da incidência cumulativa para os diferentes acontecimentos de interesse

Para obter as estimativas da função de incidência cumulativa para cada acontecimento de interesse, recorreu-se à função `cuminc` do *package* `cmprsk`. O estimador da variância usado baseia-se em (2.33) desenvolvido por Aalen.

No *output* abaixo, na Figura 3.2, a tabela `$est` indica a estimativa da função de incidência cumulativa para cada acontecimento de interesse em vários instantes de tempo de observação. Por exemplo, a primeira linha corresponde à probabilidade da estimativa da ocorrência de morte por COVID-19 aos 5, 15, 20, 50 e 100 dias. A tabela `$var` mostra a estimativa de variância aos 5, 15, 20, 50 e 100 dias, para os três acontecimentos de interesse.

```
$est
                          5           15           20           50          100
1 Morte COVID19 0,0044981295 0,0120049445 0,0136768696 0,015525276 0,01561506
1 Morte OC      0,0003811294 0,0008301919 0,0009571343 0,001214115 0,00122650
1 Recuperado    0,0000000000 0,7726661623 0,8099655727 0,983127474 0,98312747

$var
                          5           15           20           50           100
1 Morte COVID19 1,247653e-08 3,430948e-08 3,933824e-08 4,483931e-08 4,511192e-08
1 Morte OC      1,060228e-09 2,386476e-09 2,778778e-09 3,573071e-09 3,611979e-09
1 Recuperado    0,000000e+00 5,395884e-07 4,725734e-07 4,870945e-08 4,870945e-08
```

Figura 3.2: Estimativa da função de incidência cumulativa e da variância a diferentes dias para os acontecimentos de interesse.

Para uma melhor perspectiva da evolução da estimativa da função de incidência cumulativa ao longo do tempo, veja-se a Figura 3.3.

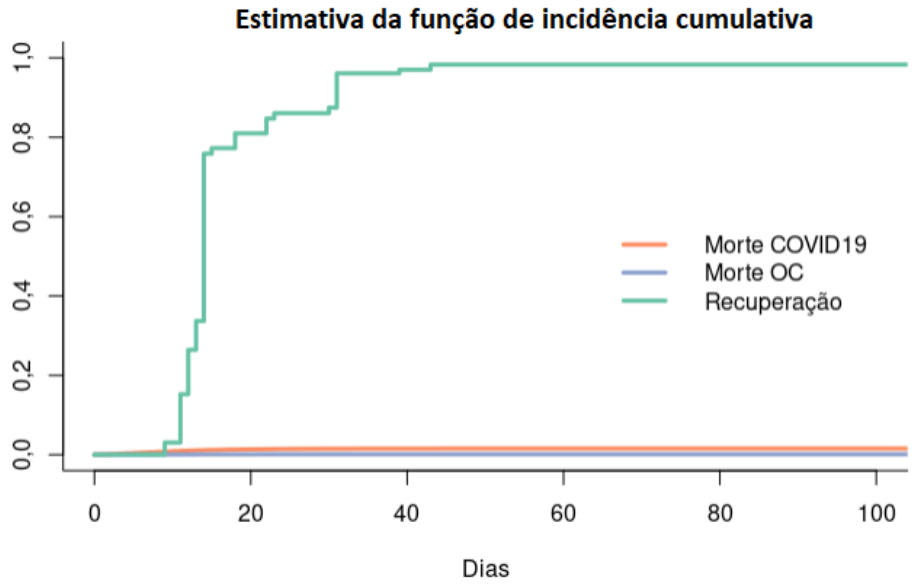


Figura 3.3: Estimativa da função de incidência cumulativa dos diferentes acontecimentos.

Tendo em conta a Tabela 3.1, sabe-se que há muito mais indivíduos cujo fim foi a recuperação e não os outros dois acontecimentos. A Figura 3.3 mostra como a estimativa da probabilidade de recuperação superior à estimativa da incidência cumulativa para qualquer um dos outros acontecimentos. O desfecho fatal devido à COVID-19 tem uma probabilidade ligeiramente superior em comparação à da morte por outra causa.

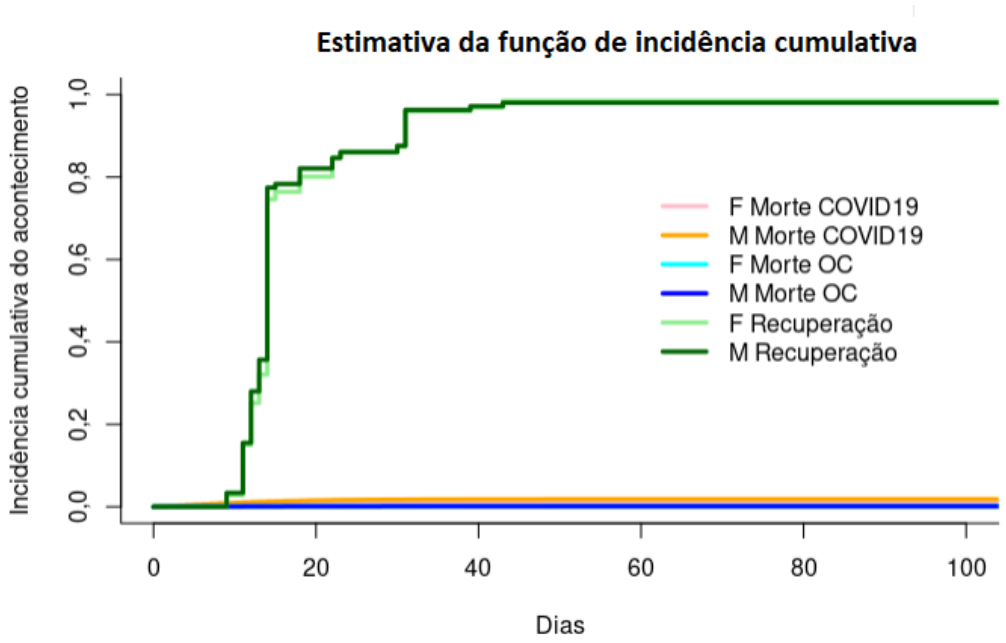


Figura 3.4: Estimativa da função de incidência cumulativa dos diferentes acontecimentos por género.
Legenda: Morte por COVID-19 no sexo feminino; Morte por COVID-19 no sexo masculino; Morte por Outra Causa no sexo feminino; Morte por Outra Causa no sexo masculino; Recuperação no sexo feminino; Recuperação no sexo masculino.

A Figura 3.4 representa a estimativa da função de incidência cumulativa para cada acontecimento e para cada género. Destaca-se o acontecimento Recuperação como o que tem maior probabilidade estimada de ocorrer nos dois sexos. Esta afirmação era esperada pois existe uma quantidade muito maior de indivíduos que recuperaram do que aqueles que morreram.

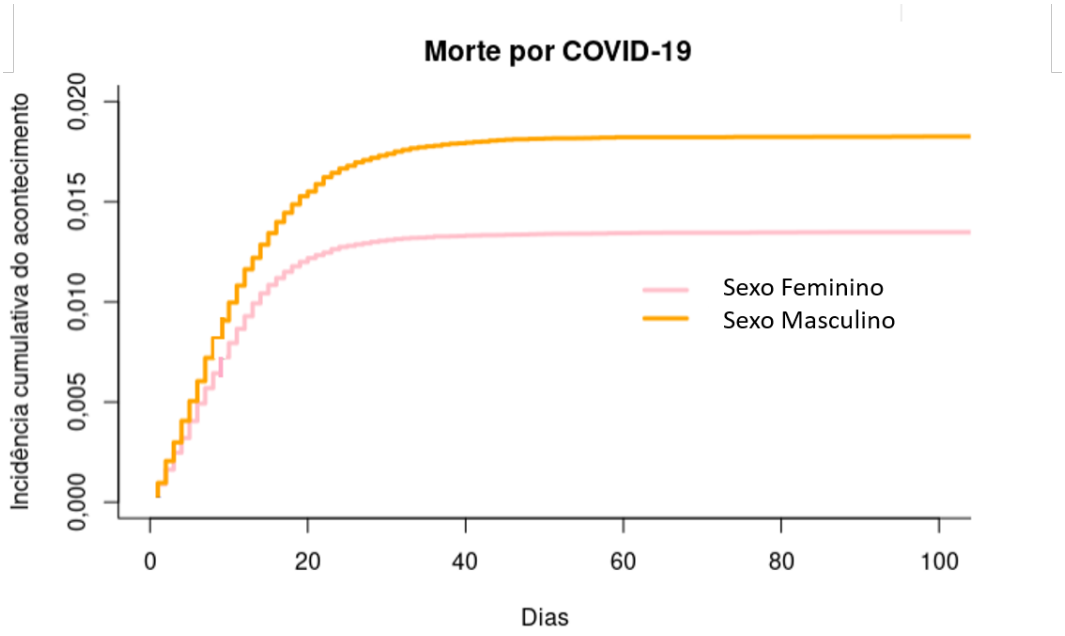


Figura 3.5: Estimativa da função de incidência cumulativa a diferentes dias para o acontecimento de interesse ”Morte por COVID-19”.

Para uma análise com mais detalhe ao acontecimento ”Morte por COVID-19”, fez-se uma ampliação

do eixo dos *yy* entre o valor 0 e o valor $0,02$. Tendo em consideração que a estimativa da função de incidência cumulativa para este acontecimento é muito baixa para ambos os sexos, a incidência cumulativa é ligeiramente superior no sexo masculino, significando que poderá existir uma diferença significativa entre os géneros e que ser do sexo masculino tem um maior risco de morte por COVID-19.

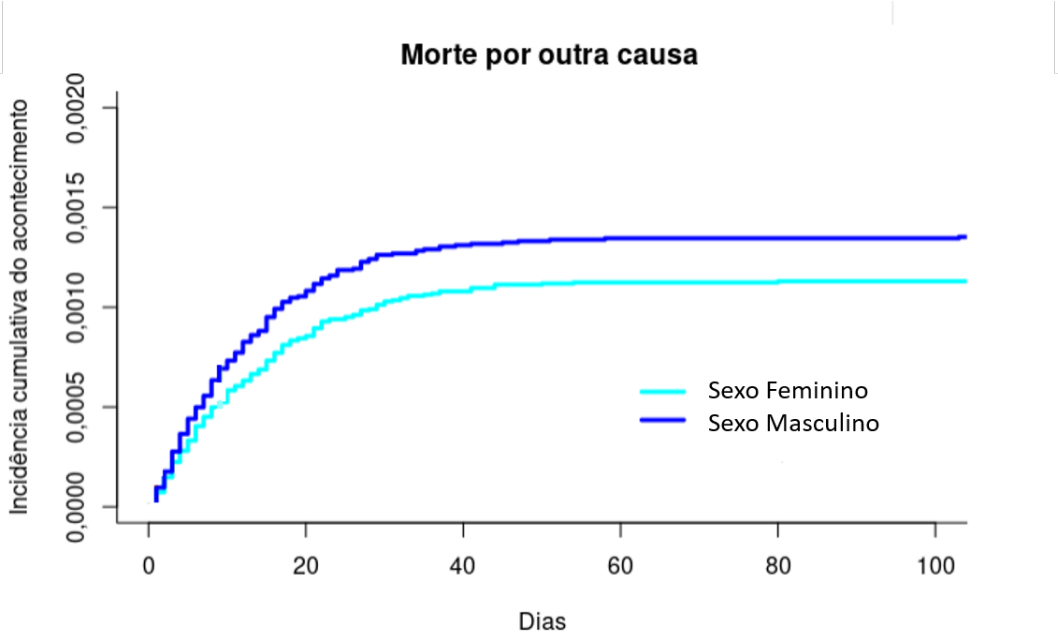


Figura 3.6: Estimativa da função de incidência cumulativa a diferentes dias para o acontecimento de interesse ”Morte por outra causa”.

Para o acontecimento ”Morte por Outra Causa”, fez-se a mesma ampliação referida anteriormente e, apesar da estimativa da função de incidência cumulativa ser extremamente baixa nos dois géneros, existe uma maior incidência cumulativa para o sexo masculino. Portanto, suspeita-se que ser do sexo masculino leva a um maior risco de morte por outra causa, comparativamente a ser do sexo feminino.

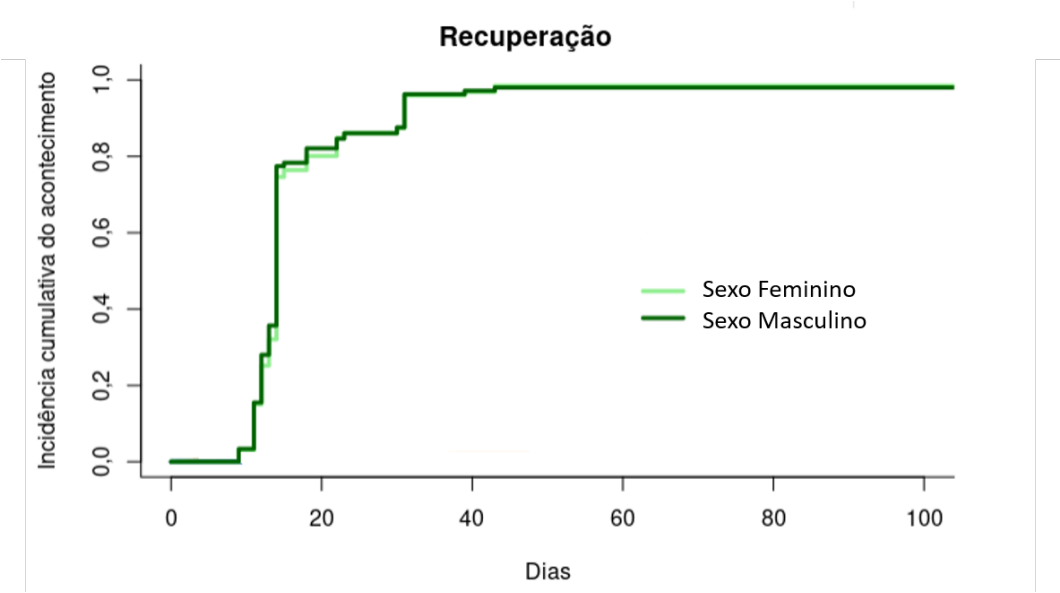


Figura 3.7: Estimativa da função de incidência cumulativa a diferentes dias para o acontecimento de interesse ”Recuperação”.

Por fim, para o acontecimento ”Recuperação”, ambos os sexos apresentam uma estimativa da função

de incidência cumulativa que toma os valores mais elevados no conjunto dos acontecimentos. Não parece existir diferença entre o sexo feminino e o sexo masculino para este acontecimento interesse.

Como referido anteriormente, na presença de riscos competitivos, o estimador de Kaplan-Meier (KM) não deve ser empregado por produzir estimativas enviesadas. O seu complemento $(1 - KM)$ apenas pode ser interpretado como a probabilidade estimada de um acontecimento de interesse ocorrer até um certo instante, num cenário hipotético em que os outros acontecimentos de interesse não pudessem ocorrer.

Calculando a estimativa de Kaplan-Meier para o acontecimento Morte por COVID-19 e comparando com a estimativa da função de incidência cumulativa, obtém-se o seguinte quadro:

Tabela 3.3: Comparação entre a estimativa de Kaplan-Meier e a estimativa da função de incidência cumulativa.

	Morte por COVID-19				
	5	**15**	**20**	**50**	**100**
Est. incidência cumulativa	0,0045	0,0120	0,0137	0,0155	0,0156
1-KM	0,0519	0,8133	0,8468	0,9998	0,9998

Como esperado, a estimativa $1 - KM$ é muitíssimo superior à estimativa da função de incidência cumulativa, principalmente a partir dos 15 dias. Verifica-se que a sobrestimação é de grande magnitude, visto que o acontecimento competitivo Recuperação é bastante frequente.

Teste de Gray

Neste estudo é importante saber se existem diferenças significativas entre os sexos, os grupos etários e as regiões, no que diz respeito à ocorrência de cada acontecimento. O teste de Gray foi utilizado para avaliar se, para cada acontecimento, existem diferenças significativas entre os vários grupos de indivíduos, definidos segundo as categorias de cada variável (`Sexo`, `Grupo etário` e `ARS`).

Tabela 3.4: Teste de Gray para comparação de grupos.

Sexo			
Acontecimento	**Valor da estatística de teste**	**Valor-p**	**Graus de liberdade**
Morte por COVID-19	138,81	≈ 0	1
Morte por outra causa	3,51	0,0592	1
Recuperação	166737,6	≈ 0	1

Grupos etários			
Acontecimento	**Valor da estatística de teste**	**Valor-p**	**Graus de liberdade**
Morte por COVID-19	28845,00	≈ 0	4
Morte por outra causa	1892,22	≈ 0	4
Recuperação	-	-	-

ARS

Acontecimento	Valor da estatística de teste	Valor-p	Graus de liberdade
Morte por COVID-19	140,39	≈ 0	6
Morte por outra causa	13	0,043	6
Recuperação	21424,94	≈ 0	6

Com base na Tabela 3.4 constata-se que para a morte por COVID-19 e para a recuperação existe uma forte evidência para afirmar que existe diferença entre os sexos, pois com base no valor-p, rejeita-se a hipótese nula, para qualquer nível de significância. Note-se que para a morte por outra causa existe uma fraca evidência para afirmar que exista diferença entre os sexos, o que pode ser justificado, em parte, por haver poucas observações relativas ao acontecimento "Morte por outras causas".

No que diz respeito aos grupos etários, há uma forte evidência para afirmar que exista diferença entre os vários grupos para os acontecimentos "Morte por COVID-19"e "Morte por outra causa". Para a recuperação, nada se conseguiu concluir devido a um erro surgido no uso do *software* que não foi possível corrigir.

Por fim, ao testar a igualdade entre ARS para cada acontecimento, concluí-se que se rejeita sempre a hipótese nula, para o nível de significância de $0,05$. Note-se que há forte evidência que existe diferença entre as regiões de saúde para os acontecimentos Morte por COVID-19 e Recuperação.

3.3 Modelação na presença de riscos competitivos

As funções de interesse estatístico são as funções de risco específica da causa e de incidência cumulativa. A primeira refere-se à taxa instantânea de ocorrência de um determinado acontecimento entre os indivíduos ainda livres de qualquer acontecimentos, enquanto que a última é a probabilidade de ocorrência de um determinado acontecimento até ao instante t.

Como referido na literatura, a análise do risco específico de um acontecimento em particular não é suficiente para a estimativa da função de incidência cumulativa desse acontecimento. Como resultado, o estimador de Kaplan-Meier é um método inadequado para estimar a incidência cumulativa na presença de acontecimentos competitivos.

E importante salientar que, para um acontecimento em particular, o efeito de uma covariável no risco específico da causa pode ser diferente do seu efeito sobre a incidência cumulativa.

Os dados de riscos competitivos são normalmente analisados usando modelos de riscos proporcionais, para o risco específico da causa e/ou risco de subdistribuição. A primeira abordagem requer que cada risco específico da causa seja modelado segundo um modelo de Cox. A segunda abordagem, também conhecida como modelo de Fine e Gray, também é um modelo de Cox, mas definido para a função de risco da subdistribuição (Fine & Gray, 1999).

Assim sendo, os dois modelos de regressão serão ajustados aos dados para dar a perceber o efeito das covariáveis recorrendo à função do risco específico da causa e à função de incidência cumulativa.

O modelo de Fine e Gray modela diretamente o efeito da covariável recorrendo à função de incidência cumulativa e permite obter a razão das funções de risco da subdistribuição. No entanto, esta razão fornece apenas a informação sobre a ordenação das curvas da função de incidência cumulativa em diferentes níveis das covariáveis, não tendo qualquer interpretação prática como os riscos proporcionais na ausência dos riscos competitivos (Fine & Gray, 1999).

O modelo de Fine e Gray pode ser ajustado com a função `crr()` do *package* `cmprsk`. Uma forma alternativa de ajustar modelos de riscos competitivos é recorrendo ao *package* `riskRegression` empregando diferentes funções de ligação entre covariáveis e a variável resposta. O pressuposto da proporcionalidade pode ser verificado testando a significância estatística dos termos de interação envolvidos no tempo até à ocorrência dos acontecimentos. Uma outra forma de verificar os pressupostos do modelo é através dos resíduos de Schoenfeld.

Para cada região, recorreu-se ao modelo de regressão de Fine e Gray e ao modelo de Cox que permitem estimar o efeito da covariável `Sexo`. Obteve-se os resultados expressos nas tabelas seguintes:

Tabela 3.5: Modelos univariável.

Modelo de Fine e Gray - Tempo até à Morte por COVID-19						
Região	**Covariável**	$\widehat{\boldsymbol{\beta}}$	**Exp**$(\widehat{\boldsymbol{\beta}})$	**SE**$(\widehat{\boldsymbol{\beta}})$	**Valor obs. da estatística de teste**	**Valor-p**
Norte	Sexo (Categoria de referência: Sexo feminino)	0,337	1,400	0,040	8,380	≈ 0
Centro		0,369	1,450	0,071	5,210	≈ 0
LVT		0,325	1,380	0,049	6,680	≈ 0
Alentejo		-0,024	0,977	0,164	-0,145	0,890
Algarve		0,183	1,200	0,268	0,683	0,490
Madeira		0,026	1,030	1,410	0,018	0,990
Açores		-1,070	0,342	0,571	-1,880	0,061

Modelo de Cox - Tempo até à Morte por COVID-19						
Região	**Covariável**	$\widehat{\boldsymbol{\beta}}$	**Exp**$(\widehat{\boldsymbol{\beta}})$	**SE**$(\widehat{\boldsymbol{\beta}})$	**Valor obs. da estatística de teste**	**Valor-p**
Norte	Sexo (Categoria de referência: Sexo feminino)	0,039	1,472	0,040	9,602	≈ 0
Centro		0,459	1,584	0,071	6,465	≈ 0
LVT		0,332	1,393	0,049	6,800	≈ 0
Alentejo		-0,029	0,972	0,165	-0,174	0,862
Algarve		0,144	1,155	0,270	0,532	0,595
Madeira		0,021	1,021	1,414	0,015	0,988
Açores		-0,954	0,385	0,578	-1,651	0,099

Recorrendo ao modelo de Fine e Gray, concluí-se que o sexo masculino está significativamente associado a um aumento da incidência de morte por COVID-19 relativamente ao sexo feminino, nas regiões Norte, Centro e Lisboa e Vale do Tejo, de 40%, 45% e 38%, respetivamente. Quanto às regiões do Alentejo e dos Açores, existe um decréscimo estimado não significativo de 2% e 66%, respetivamente, e nas regiões do Algarve e da Madeira um acréscimo não significativo de 20% e 3%,

respetivamente.

Do modelo de Cox retira-se que nas Administrações Regionais de Saúde Norte, Centro e LVT, ser homem leva a um acréscimo estimado de aproximadamente 47%, 58% e 39% no risco de morte por COVID-19, respetivamente, relativamente a ser do sexo feminino. Para as restante regiões não existe uma associação estatisticamente significativa com o Sexo, mas estima-se que ser do sexo masculino na região Alentejo e Açores leva a um decréscimo de risco de morte por COVID-19 de 3% e de 62%. Nas regiões do Algarve e da Madeira, ser homem leva a um acréscimo de 15% e 2%, respetivamente.

As conclusões das análises dos riscos específicos da causa e da incidência cumulativa estão em concordância.

3.4 Análise de resíduos

Nesta secção será apresentada a análise de resíduos referente aos modelos implementados na secção anterior.

Tanto o modelo de Cox como o modelo de Fine e Gray assumem proporcionalidade dos riscos ao longo do tempo. O pressuposto de riscos proporcionais pode ser verificado usando gráficos com base nos resíduos de Schoenfeld.

Para a covariável Sexo, a independência entre os resíduos e o tempo foi avaliada. Um gráfico que mostra um padrão não aleatório em relação ao tempo constitui evidência de violação do pressuposto de riscos proporcionais.

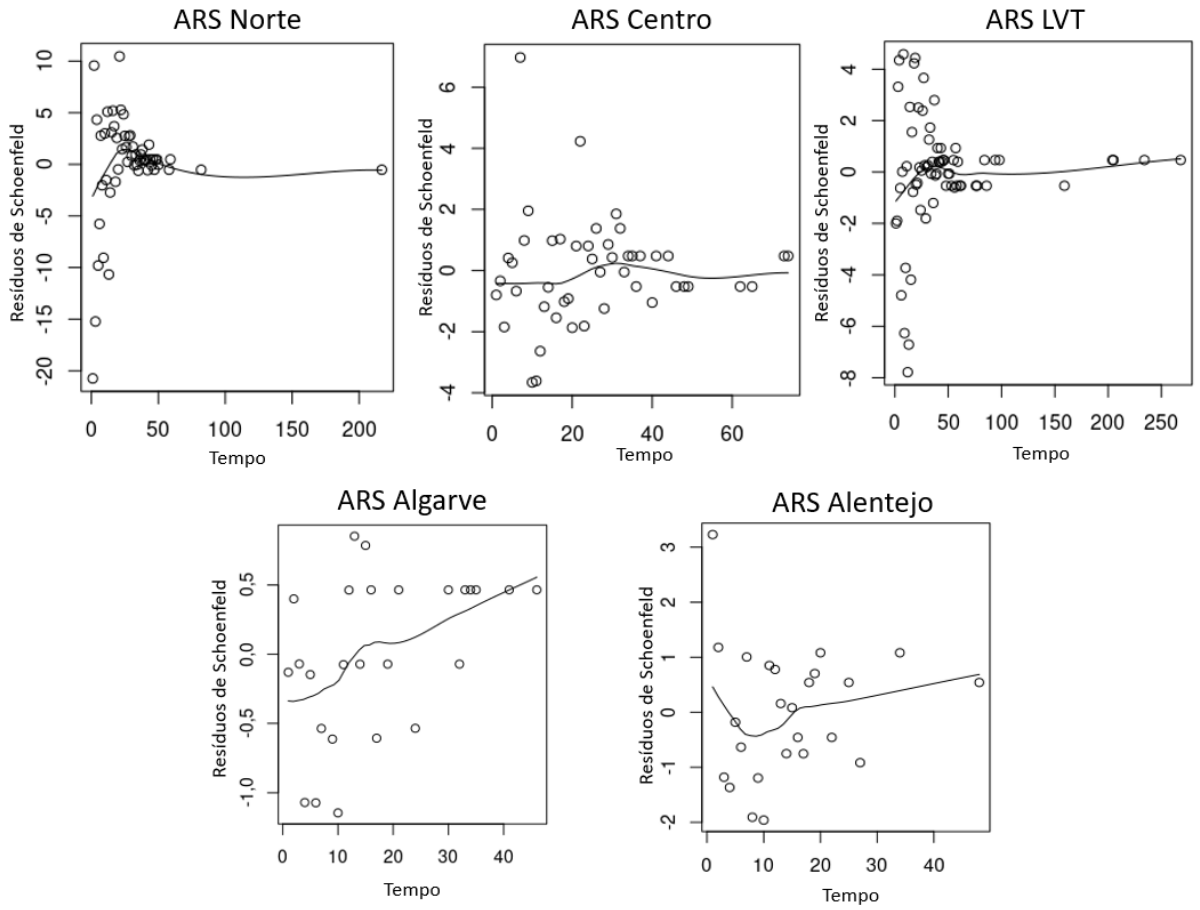


Figura 3.8: Gráficos dos resíduos de Schoenfeld do modelo de Fine e Gray com a covariável `Sexo`.

Na Figura 3.8 encontram-se representados os gráficos dos resíduos de Schoenfeld para todas as regiões, exceto para as Regiões Autónomas da Madeira e dos Açores devido ao número extremamente baixo de mortes por COVID-19, que não permitiu a construção dos gráficos.

Para as ARS Norte, Centro e LVT, a linha que representa a função suavizadora encontra-se razoavelmente horizontal e os resíduos encontram-se dispersos em torno desta de forma aleatória, sugerindo que o pressuposto de riscos proporcionais é corroborado pelos gráficos. Quanto à ARS Alentejo suspeita-se que o pressuposto de riscos proporcionais seja violado. Para a ARS Algarve parece haver evidência de não proporcionalidade das funções de risco.

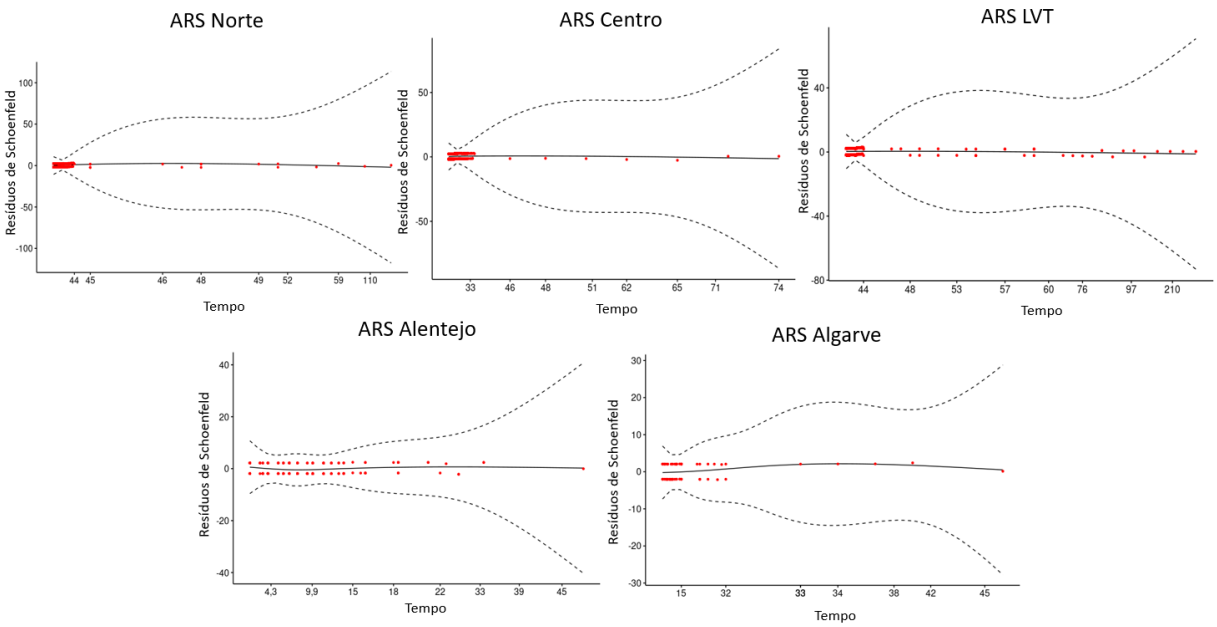


Figura 3.9: Gráficos dos resíduos de Schoenfeld do modelo de Cox com a covariável `Sexo`.

Para o modelo de Cox, os gráficos dos resíduos de Schoenfeld sugerem que a variável Sexo tem um efeito constante ao longo do tempo, isto é, o pressuposto de riscos proporcionais parece ser verificado.

Além da análise visual dos gráficos, é possível testar a hipótese de riscos proporcionais recorrendo à função `cox.zph` do *package* `survival`.

Tabela 3.6: Teste de hipóteses de proporcionalidade.

ARS	Valor obs. da est. de teste	Valor-p
Norte	2,37	0,12
Centro	1,26	0,26
LVT	5,66	0,02
Alentejo	0,46	0,50
Algarve	4,61	0,03

Sob a hipótese nula de não existir correlação entre os resíduos e o tempo (i.e, riscos proporcionais), a estatística de teste tem distribuição χ^2 com 1 grau de liberdade.

Tendo em conta os resultados do teste de hipóteses de proporcionalidade, na Tabela 3.6, constata-se que não há evidência de violação da hipótese de proporcionalidade das funções de risco exceto nas ARS LVT e Algarve.

Esta discrepância, relativamente às conclusões tiradas a partir das Figuras 3.8 e 3.9, pode resultar da existência de *outliers*.

Capítulo 4

Discussão e conclusões

Com os dados provenientes do BI SINAVE e SICO, fornecidos pela DGS, procedeu-se a um estudo semelhante a outros estudos realizados em diversas populações como, por exemplo, no Brasil, México, China e Itália. O acontecimento de interesse é a morte por COVID-19, sendo a recuperação e a morte por outra causa riscos competitivos. Nenhum dos estudos clínicos referidos, teve em consideração a presença de riscos competitivos.

Em comparação com estudos feitos noutras populações, este trabalho analisou 360 914 indivíduos, enquanto que os anteriores apenas incluíram entre 500 a 118 000 pessoas. Apesar deste trabalho apresentar uma amostra maior, as variáveis explicativas foram apenas o sexo, o grupo etário e a região de residência enquanto que nos outros trabalho tomaram em consideração o sexo, a idade, a região, mas também as comorbilidades, hábitos tabagicos, ocorrência de gravidez e outros fatores sociodemográficos.

Na amostra aqui considerada, cerca de 87% das pessoas infectadas com SARS-CoV-2 recuperaram, 1,44% morreram por COVID-19 e 0,11% morreram por outra causa, sendo que 11,62% não foi observado nenhum desses três acontecimentos, correspondendo a observações censuradas.

Se não se tivesse em consideração os riscos competitivos e se recorresse apenas a métodos clássicos da análise de sobrevivência, os resultados seriam incorretos, como evidenciado na Tabela 3.3. O facto dos diversos estudos clínicos não apresentarem uma abordagem de riscos competitivos, leva a crer que os seus resultados podem estar enviesados. A aplicação do estimador de Kaplan-Meier é baseado na premissa que a censura é não informativa, isto é, censura independente do acontecimento de interesse, o que não acontece na presença de riscos competitivos.

Considera-se que a abordagem seguida neste trabalho foi a mais correta.

Quanto ao modelo de Fine e Gray com apenas `Sexo` como covariável, conclui-se que na ARS Norte, ARS LVT e na ARS Centro, ser do sexo masculino leva a um acréscimo estimado, relativamente a ser do sexo feminino, de aproximadamente 40%, 45% e 38% na incidência de morte por COVID-19, respetivamente.

Em suma, os riscos competitivos constituem uma questão muito importante na análise de dados de sobrevivência. Os investigadores devem ter a percepção dos potenciais problemas que possam surgir associados à censura informativa. O uso da teoria de riscos competitivos requer uma pesquisa cuidada e uma selecção de métodos apropriados para a análises de dados e interpretação dos resultados.

A dimensão da amostra revelou algumas limitações dos *packages* utilizados do *software* R, uma forma de ultrapassar este problema seria recorrer a métodos de *machine learning*. Muitos algoritmos de *machine learning* têm sido adaptados para tratar de dados censurados e de outros problemas desafiantes que surgem no uso de dados reais.

Geralmente, os métodos de análise de sobrevivência podem ser classificados em duas categorias: métodos estatísticos e métodos baseado em *machine learning*. Os métodos estatísticos e os métodos de *machine learning* têm o objetivo comum de fazer previsões do tempo de sobrevivência e estimar a probabilidade de sobrevivência num certo instante. No entanto, os métodos estatísticos focam-se em caracterizar a distribuição do tempo de sobrevivência e as propriedades estatísticas dos estimadores, estimando as curvas de sobrevivência. Os métodos de *machine learning* concentram-se mais na previsão da ocorrência de certo acontecimento num determinado momento, combinando os métodos de análise de sobrevivência e de *machine learning*. Os métodos de *machine learning* têm se mostrado muito úteis em amostras de dimensões elevadas.

Uma outra abordagem alternativa à considerada neste trabalho seria recorrer a modelos multiestado, que consistem numa generalização dos modelos de análise de sobrevivência clássica. Um modelo multiestado é definido a partir de um processo estocástico que permite que os indivíduos se movam entre um número finito de estados. A abordagem dos modelos multiestado permite analisar o tempo de vida de indivíduos como um processo de mudança ou transição entre estados. Deste modo, é dada especial ênfase ao progresso da doença (generalizável a outro acontecimento) identificando os fatores que afetam as diferentes transições, proporcionando simultaneamente uma visão global da doença.

Os riscos competitivos são modelos multiestado simples, onde os acontecimentos são concebidos como transições entre estados: existe um estado inicial comum e tantos outros estados finais quantos os riscos competitivos existentes. Apenas as transições entre o estado inicial e um dos estados de riscos competitivos são considerados. Um modelo multiestado mais complexo consiste na análise das múltiplas transições possíveis entre qualquer par de estados considerados.

Numa nota final, é importante referir algumas limitações encontradas ao longo deste estudo, assim como, deixar algumas recomendações e sugestões para a estruturação e desenvolvimento de trabalhos futuros.

Uma das limitações iniciais deste estudo foi a recolha dos dados. Sendo a recolha de dados um dos passos fundamentais para um estudo rigoroso, existe uma enorme fragilidade quando os dados não são recolhidos da forma adequada. Apesar dos dados terem sido fornecidos pela Direção-Geral da Saúde e serem sujeitos a tratamento de dados por parte da Direção de Serviços de Informação e Análise, estes dados retratam uma fase de reorganização em Portugal. Como referido anteriormente, estes dados provêm do SINAVE, sistema aplicacional que se destina à desmaterialização do processo de notificação de doenças de declaração obrigatória, incluindo resistentes aos antimicrobianos (SPMS, 2020). Este sistema não tem o foco principal de servir como fonte de dados de sobrevivência, não havendo a possibilidade de existir um acompanhamento adequado dos indivíduos.

É de salientar que os médicos, delegados de saúde e restantes profissionais de saúde que inseriam dados neste sistema, encontravam-se na linha da frente ao combate da pandemia que acabara de surgir no mundo. Fatores como a exaustão por parte dos profissionais de saúde e a heterogeneidade existente na sua formação para inserir os dados de COVID-19 no sistema, levaram a uma menor qualidade na recolha dos dados.

Numa situação ideal, teria sido interessante ter acesso às comorbilidades presentes nos indivíduos que foram infetados por SARS-CoV-2, assim como, conhecer os hábitos tabágicos, o valor do IMC e a etnia. Tratar-se-ia de um estudo de maior relevância para o país pois contemplaria um número maior de potenciais fatores de risco para a morte por COVID-19. Se a recolha de dados tivesse sido feita com o intuito de se realizar uma análise de sobrevivência, provavelmente ter-se-ia resultados de maior relevância.

Adicionalmente, a dimensão da amostra revelou-se um problema para a construção dos modelos. Devido à grande quantidade de observações, não foi possível com o *software* R, ajustar um modelo de Fine e Gray sem a estratificação por Administração Regional de Saúde. Inicialmente tinha-se considerado como potencial fator de risco a área de residência do utente, considerando a variável `ARS` como uma covariável. Para contornar esta adversidade, construíram-se os modelos estratificados por ARS usando apenas a variável `Sexo` como variável explicativa.

Uma última recomendação, mas não menos importante, é estender o período de observação até março de 2021, para haver um ano de dados de sobrevivência ou até junho do mesmo ano, incluindo os meses onde houve mais casos de infeção por SARS-CoV-2 e mais mortes por COVID-19.

Bibliografia

Austin, P. C., Lee, D. S. & Fine, J. P. (2016). Introduction to the Analysis of Survival Data in the Presence of Competing Risks. *Circulation*, *133*(6), 601–609. https://doi.org/10.1161/circulationaha.115.017719

Benichou, J. & Gail, M. H. (1990). Estimates of Absolute Cause-Specific Risk in Cohort Studies. *Biometrics*, *46*(3), 813. https://doi.org/10.2307/2532098

Collett, D. (2015). *Modelling survival data in medical research*. CRC Press, Taylor & Francis Group.

Cox, D. R. (1972). Regression Models and Life-Tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, *34*(2), 187–202. https://doi.org/https://doi.org/10.1111/j.2517-6161.1972.tb00899.x

DGS. (2020). Plano de Vacinação contra a COVID-19. https://covid19.min-saude.pt/wp-content/uploads/2020/12/PLANO-VACINA%C3%87%C3%83O_20201203.pdf

Fine, J. P. & Gray, R. J. (1999). A Proportional Hazards Model for the Subdistribution of a Competing Risk. *Journal of the American Statistical Association*, *94*(446), 496–509. https://doi.org/10.1080/01621459.1999.10474144

Galvão, M. H. & Roncalli, A. G. (2020). Fatores associados a maior risco de ocorrência de óbito por COVID-19: análise de sobrevivência a partir de casos confirmados. *Revista Brasileira de Epidemiologia*. https://doi.org/10.1590/scielopreprints.1175

Gray, R. J. (1988). A Class of *K*-Sample Tests for Comparing the Cumulative Incidence of a Competing Risk. *The Annals of Statistics*, *16*(3). https://doi.org/10.1214/aos/1176350951

Kalbfleisch, J. D. & Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*. John Wiley & Sons, Inc. https://doi.org/10.1002/9781118032985

Klein, J. P., van Houwelingen, H. C., Ibrahim, J. G. & Scheike, T. H. (2016). *Handbook of Survival Analysis*. Taylor Francis Ltd. https://www.ebook.de/de/product/22287974/handbook_of_survival_analysis.html

Lau, B., Cole, S. R. & Gange, S. J. (2009). Competing Risk Regression Models for Epidemiologic Data. *American Journal of Epidemiology*, *170*(2), 244–256. https://doi.org/10.1093/aje/kwp107

Li, X., Xu, S., Yu, M., Wang, K., Tao, Y., Zhou, Y., Shi, J., Zhou, M., Wu, B., Yang, Z., Zhang, C., Yue, J., Zhang, Z., Renz, H., Liu, X., Xie, J., Xie, M. & Zhao, J. (2020). Risk factors for severity and mortality in adult COVID-19 inpatients in Wuhan. *Journal of Allergy and Clinical Immunology*, *146*(1), 110–118. https://doi.org/10.1016/j.jaci.2020.04.006

Malagón-Rojas, J., Ibáñez, E., B, E. L. P., Toloza-Perez, Y. G., Álvarez, S. & Mercado, M. (2021). Analysis of COVID-19 Mortality and Survival in Colombia: A prospective Cohort Study. *Asociación Colombiana de Infectología*, *25*(3), 176. https://doi.org/10.22354/in.v25i3.943

Pepe, M. S. (1991). Inference for Events with Dependent Risks in Multiple Endpoint Studies. *Journal of the American Statistical Association*, *86*(415), 770–778. https://doi.org/10.1080/01621459.1991.10475108

Pepe, M. S. & Mori, M. (1993). Kaplan—Meier, marginal or conditional probability curves in summarizing competing risks failure time data? *Statistcs in Medicine*, *12*(8), 737–751. https://doi.org/10.1002/sim.4780120803

Pintilie, M. (2006, setembro 28). *Competing Risks: A Practical Perspective*. John Wiley & Sons. https://www.ebook.de/de/product/6542612/pintilie_competing_risks.html

Putter, H., Fiocco, M. & Geskus, R. B. (2007). Tutorial in biostatistics: competing risks and multi-state models. *Statistics in Medicine*, *26*(11), 2389–2430. https://doi.org/10.1002/sim.2712

Rocha, C. & Papoila, A. L. (2009). *Análise de Sobrevivência*. Edições SPE.

Salinas-Escudero, G., Carrillo-Vega, M. F., Granados-Garcıa, V., Martınez-Valverde, S., Toledano-Toledano, F. & Garduño-Espinosa, J. (2020). A survival analysis of COVID-19 in the Mexican population. *BMC Public Health*, *20*(1). https://doi.org/10.1186/s12889-020-09721-2

Satagopan, J. M., Ben-Porat, L., Berwick, M., Robson, M., Kutler, D. & Auerbach, A. D. (2004). A note on competing risks in survival data analysis. *British Journal of Cancer*, *91*(7), 1229–1235. https://doi.org/10.1038/sj.bjc.6602102

Schoenfeld, D. (1982). Partial residuals for the proportional hazards regression model. *Biometrika*, *69*(1), 239–241. https://doi.org/10.1093/biomet/69.1.239

SPMS. (2020). SINAVE — Sistema Nacional de Vigilância Epidemiológica. https://www.spms.min-saude.pt/2020/07/sinave-2/

WHO. (2020). Coronavirus diseade (COVID-19). https://www.who.int/health-topics/coronavirus#tab=tab_1

Wolbers, M., Koller, M. T., Stel, V. S., Schaer, B., Jager, K. J., Leffondré, K. & Heinze, G. (2014). Competing risks analyses: objectives and approaches. *European heart journal*. https://doi.org/10.1093/eurheartj/ehu131