Universidade do Minho
Escola de Ciências

Sara Raquel Melo Magalhães

Modelos conjuntos para dados longitudinais no cancro da mama e tempo até recidiva

dezembro de 2021

Modelos conjuntos para dados longitudinais no cancro da mama e tempo até recidiva

Sara Raquel Melo Magalhães

UMinho | 2021

Universidade do Minho
Escola de Ciências

Sara Raquel Melo Magalhães

Modelos conjuntos para dados longitudinais no cancro da mama e tempo até recidiva

Dissertação de Mestrado
Mestrado em Estatística

Trabalho efetuado sob a orientação da
Professora Doutora Inês Pereira Silva Cunha Sousa

dezembro de 2021

Direitos de autor e condições de utilização do trabalho por terceiros

Agradecimentos

Com a finalização desta etapa muito importante da minha vida profissional e pessoal, não posso deixar de agradecer àqueles que me acompanharam neste percurso. Em primeiro lugar, agradeço aos meus pais porque, sem eles, nada seria possível. Agradeço por terem sonhado mais alto que eu, por me terem incentivado a lutar por um futuro mais promissor e por me apoiarem incondicionalmente ao longo destes anos. Ao Tiago, agradeço por sempre me incentivar a seguir os meus objetivos, por acreditar em mim e por me acompanhar ao longo deste percurso. À minha orientadora, Professora Inês Sousa, agradeço pela ajuda prestada, por todos os seus conselhos e por acreditar no meu trabalho ao me acolher no seu projeto de investigação. A todos os professores com os quais tive oportunidade de aprender neste mestrado, agradeço por todo o conhecimento partilhado. Aos meus colegas, o meu sincero agradecimento pelo voto de confiança para os representar ao longo deste percurso. Em particular, agradeço à minha colega Inês Fortes por toda a ajuda e disponibilidade.

Este trabalho teve o apoio do projeto 028248/SAICT/2017 financiado pelo Programa Operacional Competitividade e Internacionalização (COMPETE2020) na sua componente de Fundo Europeu de Desenvolvimento Regional (FEDER) e pela Fundação para a Ciência e a Tecnologia, I.P. (FCT, I.P.) na sua componente OE.

Declaração de Integridade

Declaro ter atuado com integridade na elaboração do presente trabalho académico e confirmo que não recorri à prática de plágio nem a qualquer forma de utilização indevida ou falsificação de informações ou resultados em nenhuma das etapas conducente à sua elaboração. Mais declaro que conheço e que respeitei o Código de Conduta Ética da Universidade do Minho.

Resumo

Título: Modelos conjuntos para dados longitudinais no cancro da mama e tempo até recidiva

Modelos conjuntos para dados longitudinais e tempo até ao evento de interesse são usados em bases de dados onde o marcador longitudinal está associado com o evento de interesse. Para este estudo consideram-se dois marcadores tumorais utilizados no diagnóstico e acompanhamento deste tipo de cancro: o Antigénio Carcino-Embrionário (CEA) e o Antigénio cancerígeno 15-3 (CA15-3). O evento de interesse considerado é a recidiva do cancro após diagnóstico de cancro da mama. É expectável que o tempo até à recidiva esteja associado com a evolução da doença, traduzida na evolução dos marcadores tumorais.

Outro estudo foi já desenvolvido usando como evento de interesse a morte do paciente, o que mostrou haver uma forte associação entre os dois processos. Agora pretende-se verificar se o mesmo tipo de associação está presente para a recidiva. Existem já identificados alguns riscos para o desenvolvimento do cancro da mama e o objetivo é perceber quais destes estão relacionados com a recidiva da doença.

No presente estudo analisou-se uma base de dados com pacientes diagnosticados com cancro da mama, entre 2008 e 2012, acompanhados no Hospital de Braga. Consideraram-se medidas repetidas dos dois marcadores tumorais, a cada 6 meses aproximadamente e o tempo desde o diagnóstico até recidiva do cancro.

Como metodologia estatística utilizaram-se modelos longitudinais com efeitos aleatórios, análise de sobrevivência fazendo-se estimação de curvas de Kaplan-Meier e modelos de regressão de Cox e modelos de efeitos partilhados para a análise de modelos conjuntos.

Palavras-chave: Cancro da mama, longitudinal, modelos conjuntos, recidiva, tempo-até-evento.

Abstract

Title: Joint models for longitudinal data on breast cancer and time to relapse

Joint models for longitudinal and time-to-event data are used in databases where the longitudinal marker is associated with the event of interest. In this study, two tumor markers were used in the diagnosis and monitoring of this type of cancer: the Carcinoembryonic antigen (CAE) and the Carcinoma Antigen 15-3 (CA15-3). The event of interest considered is cancer recurrence after breast cancer diagnosis. It is expected that the time until recurrence is associated with the evolution of the disease, translated into the evolution of the tumor markers.

Another study has already been developed using the patient's death as the event of interest, which showed that there is a strong association between the two processes. Now we intend to verify if the same type of association is present for the recurrence. There are already some risks identifed for the development of breast cancer and the objective of this study is to understand which of those are related to the recurrence of breast cancer.

An extensive analysis of a database with patients diagnosed with breast cancer between 2008 and 2012 was carried out at Braga's Hospital. Repeated measurements of the two tumor markers approximately every 6 months and the time from the last diagnosis to the cancer recurrence were used.

The statistical methodology used was longitudinal models with random effects, survival analysis by estimating Kaplan-Meier curves and Cox regression models, and shared random-effects models for the analysis of joint models.

Keywords: Breast cancer, joint models, longitudinal, recurrence, time-to-event.

Conteúdo

Lista de Abreviaturas

v.a Variável aleatória

MCAR Omissão completamente aleatório *(Missing Completely At Random)*

MAR Omissão aleatório *(Missing At Random)*

MNAR Omissão não aleatória *(Missing Not At Random)*

CEA Antigénio Carcino-Embrionário *(Carcinoembryonic Antigen)*

CA15-3 Antigénio cancerígeno 15-3 *(Carcinoma Antigen 15-3)*

HR *Hazard ratio*

Base de dados TT Base de dados com todos os pacientes e todas as observações

Base de dados TR Base de dados com todos os pacientes e observações até recidiva

Base de dados RT Base de dados com pacientes com recidiva e todas as observações

Base de dados RR Base de dados com pacientes com recidiva e observações até recidiva

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

O principal objetivo deste trabalho consiste no estudo da recidiva do cancro da mama na Unidade de Senologia do Hospital de Braga. Como referido no site da Liga Portugesa Contra o Cancro[1], considera-se que há recidiva do cancro da mama quando após o tratamento o cancro reaparece. Com os sucessivos avanços no tratamento do cancro da mama, há um número crescente de pessoas que conseguem sobreviver à doença. Estes pacientes são continuamente seguidos para que se consiga detetar precocemente uma recidiva, se esta acontecer (Vasconcelos et. al., 2017).

Para a base de dados disponível, que contém informação das pacientes com diagnóstico de cancro da mama entre os anos de 2008 e 2012, no Hospital de Braga, e de todos os pacientes em acompanhamento à data de 1 de janeiro de 2008, consideram-se dois marcadores tumorais utilizados no diagnóstico e acompanhamento deste tipo de cancro: o CEA e o CA15-3, e a recidiva do cancro após diagnóstico de cancro da mama como evento de interesse. Os marcadores tumorais são variáveis longitudinais e estão disponíveis medidas repetidas a cada 6 meses, aproximadamente.

Para a análise dos marcadores tumorais começou-se por considerar diferentes bases de dados e analisou-se as diferenças encontradas nas progressões das pacientes. Esta análise foi efetuada uma vez que, como o evento de interesse é a recidiva, existem observações após o evento e esta análise permitiu inferir sobre as diferenças nas estimações dos modelos tendo em conta as diferentes bases de dados. Por fim, tendo em conta todas as observações, apresenta-se um modelo de efeitos aleatórios com a estrutura de correlação que melhor representa a variabilidade dentro dos sujeitos, para cada marcador tumoral. Na análise do tempo-até-evento foram efetuadas curvas de sobrevivência de Kaplan-Meier (1958) e é apresentado um modelo de riscos proporcionais de Cox (1972).

Um estudo anteriormente desenvolvido, referente à mesma base de dados, que considerou como evento de interesse a morte do paciente mostrou existir uma forte associação entre os processos longitudinais e de sobrevivência (Borges, 2015). Pretende-se verificar

[1] Consultado a 5 de Maio de 2021, retirado de *https://www.ligacontracancro.pt/cancro-da-mama/*

se o mesmo tipo de associação está presente quando se considera a recidiva do cancro como evento de interesse. Assim, após análises de sobrevivência e longitudinais separadas, foi realizada uma modelação conjunta destes dois processos para inferir sobre a sua associação, adotando a metodologia de efeitos aleatórios partilhados.

Esta dissertação está dividida em cinco partes. Após esta introdução, apresenta-se, no Capítulo 2, a problemática do cancro da mama em Portugal e uma breve apresentação do Hospital de Braga. No Capítulo 3 são apresentados os modelos estatísticos utilizadas neste estudo. Começa-se por apresentar a análise longitudinal, seguida da análise de sobrevivência e por fim a análise de modelos conjuntos. Os resultados das diferentes análises efetuadas, análise exploratória da base de dados, análises longitudinais dos dois marcadores, análise de sobrevivência do evento de interesse e análise conjunta dos dois processos, são apresentados no Capítulo 4. As principais conclusões e trabalho futuro são exibidos no Capítulo 5.

Capítulo 2

Cancro da mama em Portugal

O cancro da mama é o tumor maligno mais frequente na mulher, sendo um problema de saúde pública que afeta tanto os países desenvolvidos como os países em vias de desenvolvimento (Rodrigues, 2011). Num processo normal, as células crescem e dividem-se conforme o necessário e quando as células envelhecem ou são anormais elas morrem. Este tipo de cancro desenvolve-se quando as células começam a crescer descontroladamente na mama havendo produção de novas células sem que as velhas ou anormais morram quando devem. A maioria dos cancros da mama começam nos dutos que levam o leite ao mamilo, mas há cancros que começam nas glândulas que produzem o leite materno e outros tipos, que são menos comuns, como o tumor filodes e angiossarcoma. Salienta-se que nem todos os cancros tem um nódulo associado, no entanto os mais comuns são os carcinomas ductal in situ e os carcinomas invasivos. Para informação mais detalhada recomenda-se o site da *American Cancer Society*[1] de onde foi retirada esta informação. Segundo Rodrigues (2011) a incidência e a mortalidade são os indicadores habitualmente usados para quantificar a magnitude do problema oncológico. A incidência disponibiliza uma aproximação do risco médio de desenvolver um cancro, contabilizando o número de novos casos ocorridos numa determinada população num determinado período de tempo (geralmente 1 ano) ou também pode ser expressa como uma taxa que relaciona o valor anteriormente obtido por 100000 habitantes/ano. Conforme se verifica no gráfico que apresenta a evolução das principais patologias oncológicas entre 2006 e 2010 em Portugal, disponível no *Relatório Doenças Oncológicas em Números 2015* (Figura 2.1), a incidência do cancro da mama em Portugal tem aumentado desde 2006, sendo o cancro com maior incidência em 2010 na população feminina. Este facto também pode ser confirmado na Tabela 2.1 que apresenta o TOP 10 cancro em Portugal em 2010 disponível no *Relatório Doenças Oncológicas em Números 2015*.

[1] Consultado a 7 de Maio de 2021, retirado de *https://www.cancer.org/cancer/breast-cancer.html*

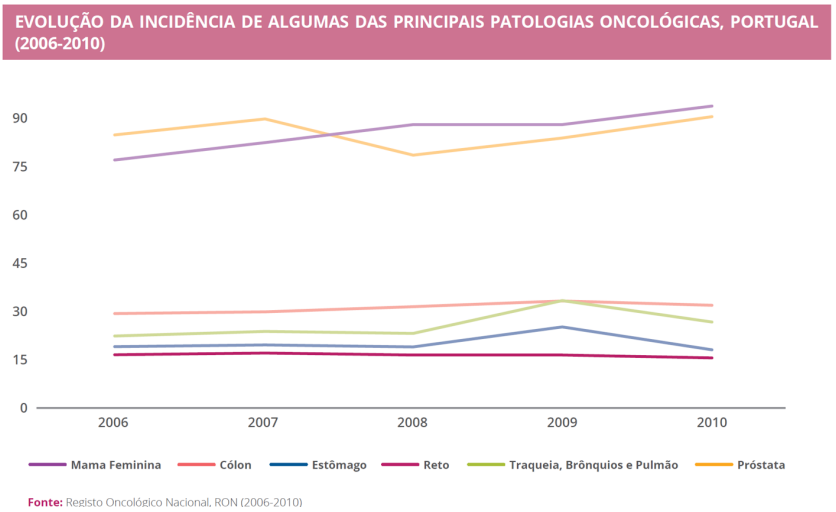

Figura 2.1: Evolução da incidência de algumas das principais patologias oncológicas, Portugal (2006-2010)

Tabela 2.1: Taxa de incidência de tumores malignos no sexo feminino (2010), RORENO, RON, 2010

Taxa de incidência de tumores malignos (100000 habitantes), no sexo feminino (2010)		
	Taxa bruta	**Taxa pad. (pop. Eur.)**
Mama	118,5	93,2
Cólon	39,0	24,2
Glândula tiroideia	23,8	21,5
Estômago	21,3	13,1
Corpo do útero	17,8	12,4
Reto	16,3	10,4
Traqueia, brônquios e pulmão	15,8	11,0
Linfoma não Hodgkin	15,3	10,8
Colo do útero	13,5	11,3
Melanoma maligno da pele	9,1	6,9
Total	382,7	279,6

A mortalidade por cancro expressa-se da mesma forma, mas contabilizando-se o número de óbitos (por tipo) de cancro. Este indicador é um produto entre a incidência e a letalidade (proporção de doentes que morreram devido à doença) e é muito influenciado pelas taxas de incidência, pela acessibilidade aos cuidados de saúde e pela eficiência das intervenções terapêuticas, além da qualidade dos certificados de óbito (Rodrigues, 2011). A taxa de mortalidade por cancro da mama em Portugal tem vindo a aumentar entre os anos de 2008 e 2012 como se verifica na Tabela 2.2 disponível no *Relatório Doenças Oncológicas em Números 2014*.

Tabela 2.2: Taxa de mortalidade do cancro da mama 2008-2012, INE, IP, 2014

Tumor maligno da mama feminino					
	2008	**2009**	**2010**	**2011**	**2012**
Número de óbitos	1504	1538	1571	1546	1663
Taxa de mortalidade	28,8	29,4	29,9	29,5	31,8
Taxa de mortalidade padronizada	19,2	19,6	19,4	18,6	19,6

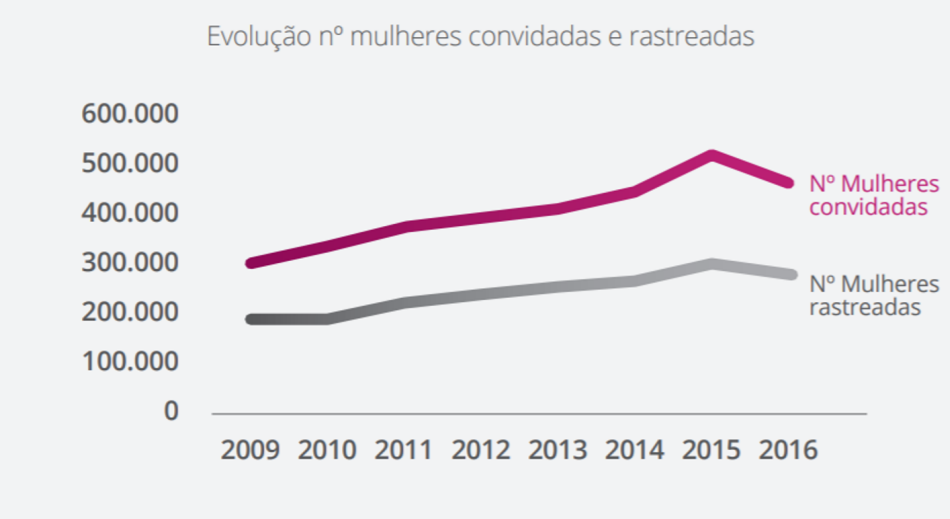

Figura 2.2: Rastreio em Portugal, 2009-2016, Programa nacional para as doenças oncológicas 2017

O rastreio, definido como a aplicação de um teste para identificar doença assintomática precocemente, tem um importante papel na redução da mortalidade por cancro da mama (Rodrigues, 2011). Em Portugal, de acordo com o gráfico da evolução do número de mulheres convidadas para rastreio e de mulheres rastreadas disponível no *Relatório Programa Nacional Para as Doenças Oncológicas 2017* (Figura 2.2), verifica-se um aumento destes números entre 2009 e 2015. Em 2016 estes números sofreram um ligeiro decréscimo.

A recidiva do cancro da mama está presente quando o cancro reaparece após o tratamento como mencionado no site da Liga Portuguesa Contra o Cancro[2]. Apesar de em Portugal a taxa de mortalidade ter aumentado entre 2008 e 2012, o número de pacientes em seguimento tem aumentado dada a diminuição da mortalidade provocada pela doença resultante dos avanços dos tratamentos disponíveis. "O risco de recidiva pode e deve ser estratificado em função das caraterísticas biológicas e moleculares do tumor, da paciente e da própria mama". Vasconcelos et. al. (2017) referem que pacientes com subtipos moleculares triplo negativo e HER2+, cancros com alto grau nuclear, mulheres jovens com idade inferior a 40 anos, risco genético associado, antecedentes de radioterapia, e em mulheres com padrões mamários densos que apresentam menor sensibilidade mamográfica existe um risco mais elevado de recidiva da doença. Dado o risco associado à recidiva desta doença e à pouca informação disponível sobre a recidiva do cancro da mama em Portugal é muito importante procurar-se identificar quais riscos estão associados à recidiva, para se poder adequar o seguimento e os tratamentos com o objetivo de diminuir o número de recidivas nas pacientes diagnosticadas com a doença em Portugal.

[2]Consultado a 5 de Maio de 2021, retirado de *https://www.ligacontracancro.pt/cancro-da-mama/*

2.1 Hospital de Braga

O Hospital de Braga[3] está localizado na cidade de Braga a norte de Portugal. Este hospital presta cuidados de saúde a cerca de 1,2 milhões de pacientes dos distritos de Braga e Viana do Castelo e concilia nas suas instalações as unidades de assistência médica, investigação e ensino universitário. Tem conquistado o primeiro lugar no grupo de melhores hospitais de média/grande dimensão do Serviço Nacional de Saúde nos prémios "TOP 5 - A Excelência dos Hospitais", promovido pela IASIST – empresa multinacional de benchmarking hospitalar, desde 2015. Em 2008 foi criada a Unidade de Senologia que é o ramo da medicina que estuda a anatomia, a fisiologia e as patologias da mama, de onde foi recolhida toda a informação que será utilizada no decorrer deste trabalho. O Hospital de Braga teve novas instalações em Maio de 2011.

[3]Consultado a 18 de Maio de 2021, retirado de *https://www.hospitaldebraga.pt/hospital/sobre-nos*

Capítulo 3

Modelos estatísticos

Neste capítulo apresentam-se os modelos estatísticos utilizados nas análises longitudinais, de sobrevivência e de modelos conjuntos. Começa-se por apresentar os modelos para a análise longitudinal explicando-se a abordagem utilizada, os mecanismos de dados omissos, o estudo do variograma e a análise de diagnóstico para validação dos pressupostos do modelo. Em seguida, na análise de sobrevivência, apresenta-se a estimação não paramétrica da função de sobrevivência, a comparação de curvas de sobrevivência, o modelo de riscos proporcionais de Cox e a análise de diagnóstico. Por fim, apresentam-se as diferentens abordagens de modelos conjuntos, os *packages* utilizados para a implementação dos modelos no *software* R bem como a análise de diagnóstico.

3.1 Análise longitudinal

Dados longitudinais são caracterizados por medidas repetidas ao longo do tempo de uma dada variável resposta em vários indivíduos. É comum que nos estudos de medidas repetidas se assuma que os diferentes indivíduos envolvidos no estudo são independentes. No entanto, assumir-se que há independência entre as medidas de um mesmo sujeito não é adequado dado que as medidas ao longo do tempo de um indivíduo tendem a estar correlacionadas. Ademais, há erros de medição associados às medições de diferentes sujeitos e dentro de um mesmo indivíduo (Sousa, 2011).

Neste estudo o interesse reside na análise dos valores de dois marcadores tumorais: CEA e CA15-3, onde são aplicados modelos longitudinais testando-se diferentes estruturas de correlação.

Seja Y_{ij} uma variável resposta medida no sujeito $i = 1, \ldots, n$ no tempo t_{ij} onde $j = 1 \ldots, m_i$. Considera-se um conjunto de p variáveis explicativas dadas pelo vetor $\boldsymbol{x_{ij}}$ de dimensão p, que podem ser medidas ao longo do tempo (variáveis dependentes do tempo), ou apenas em *baseline*, neste estudo medidas apenas no momento do diagnóstico. O vetor $\boldsymbol{Y_i} = (Y_{i1}, \ldots, Y_{im_i})$ representa o conjunto completo de medidas repetidas para o sujeito i, que tem média $E[\boldsymbol{Y_i}] = \boldsymbol{\mu_i}$ e matriz de variância-covariância dado por $Var(\boldsymbol{Y_i}) = V_i$ com

dimensão $(m_i \times m_i)$, onde cada elemento (j,k) da matriz tem covariância $Cov(Y_{ij}, Y_{ik}) = v_{ij_k}$ e $Var(Y_{ij}) = v_{ij}$, para $j = k$.

Na abordagem mais comum para dados longitudinais a independência entre os sujeitos i é assumida e cada medida obtida é considerada uma realização de uma variável aleatória com distribuição gaussiana. O modelo linear é baseado na regressão de variáveis explicativas,

$$Y_{ij} = \mu_i(t_{ij}) + \epsilon_{ij}$$

As diferentes estruturas de correlação dos erros ϵ_{ij} resultam em modelos longitudinais diferentes, o que será explicado mais à frente. A notação $\boldsymbol{Y} = (\boldsymbol{Y_1}, \ldots, \boldsymbol{Y_n})$ representa todas as medidas de todos os sujeitos para o conjunto de dados total $N = \sum_{i=1}^{n} m_i$. Assim, o modelo linear para as medidas longitudinais é,

$$\boldsymbol{Y} \sim MVN(\boldsymbol{X\beta}, \boldsymbol{V}(\psi)),$$

onde $\boldsymbol{X}$ é uma matriz de desenho de dimensão $(N \times (p+1))$ e $\boldsymbol{\beta}$ é o vetor de coeficientes das variáveis explicativas de dimensão $p+1$. Como se assume independência entre os indivíduos, $\boldsymbol{V}$ é uma matriz bloco-diagonal de dimensão $(N \times N)$ e parâmetro ψ, onde cada matriz diagonal $\boldsymbol{V_i}$ é a matriz de variância-covariância do sujeito i.

Quando se assume uma estrutura de correlação para os erros ϵ_{ij}, separando o erro de medição puro da variabilidade entre e dentro dos indivíduos, aplica-se a ideia base dos modelos lineares de efeitos mistos. Este modelo é geralmente definido como,

$$Y_{ij} = \mu_i(t_{ij}) + \omega_i(t_{ij}) + Z_{ij}$$

onde $\omega_i(t_{ij})$ é um processo aleatório não observado e Z_{ij} são realizações de uma variável aleatória gaussiana com média zero e variância τ^2, que representa o erro de medição puro, ou seja, a variabilidade não explicada. Diggle, Heagerty, Liang, e Zeger (2002) sugeriram, decompor, em duas componentes, o processo aleatório não observado $\omega_i(t_{ij})$,

$$\omega_i(t_{ij}) = \boldsymbol{d'}_{ij}\boldsymbol{U}_i + W_i(t_{ij})$$

sendo $\boldsymbol{d}_{ij}$ o vetor de variáveis explicativas de dimensão r para os efeitos aleatórios $\boldsymbol{U}_i$ que são n realizações independentes de uma variável aleatória gaussiana multivariada de dimensão r com média zero e matriz de variância-covariância G. $W_i(t_{ij})$ são n realizações independentes de um processo estocástico gaussiano com média zero e variância σ^2, representando a variabilidade dentro do indivíduo, ou seja, a variância do processo ao longo do tempo, e função de correlação $\rho(u)$, com $u = |t_{ik} - t_{ij}|$. Como já havia sido anteriormente referido, diferentes estruturas de correlação resultam em modelos longitudinais diferentes. Assim, considerando uma estrutura de correlação exponencial dentro dos indivíduos,

$$\rho(u) = \exp(-\tfrac{1}{\phi}\ |u|)$$

por sua vez, uma estrutura de correlação gaussiano tem a forma,

$$\rho(u) = \exp(-\tfrac{1}{\phi}\ u^2)$$

onde ϕ é o parâmetro *range*, valor a partir do qual a correlação estabiliza.

3.1.1 Dados omissos

Geralmente, quando se trabalha com dados reais surgem dados omissos e é muito relevante reconhecer o mecanismo que está por detrás da falta de medidas da variável resposta. Quando, como neste trabalho, se analisam dados não balanceados, ou seja, quando a variável resposta não é medida nos mesmos tempos em todos os indivíduos, é difícil identificar quando de facto se perde uma medida. No entanto, deve-se incorporar, se necessário, o mecanismo de dados omissos tendo em conta as diferentes classificações dos mesmos. Seguindo a notação de Diggle et. al (2002), considera-se $Y^* = (Y_0, Y_m)$ o conjunto completo de medições sem valores ausentes onde Y_0 são as medidas realmente obtidas (observadas) e Y_m são as medidas que estariam disponíveis se não houvesse perda de dados. Considera-se, também, o conjunto de variáveis aleatórias indicadoras R, que apresenta quais elementos de Y^* pertencem a Y_0 e quais pertencem a Y_m. A distribuição de probabilidade de R condicional a Y^* é definida pelo modelo probabilístico para o mecanismo de dados omissos.

Segundo Little e Rubin (2000), o mecanismo de dados omissos pode ser classificado em três diferentes categorias:

- **MCAR** (*Missing Completely At Random*) Omissão completamente aleatória se R é independente de Y_0 e Y_m. Por exemplo, se o paciente se esquece que tinha consulta marcada e não aparece.

- **MAR** (*Missing At Random*) Omissão aleatória se R é independente de Y_m. Por exemplo, se o médico aconselha a saída do doente do estudo com base em medidas longitudinais observadas anteriormente.

- **MNAR** (*Missing Not At Random*) Omissão não aleatória se R depende de Y_m. Por exemplo, quando um paciente sai do estudo porque não se encontra bem no dia da consulta, e o perfil longitudinal está relacionado com a doença, inclusive as medidas que teriam sido observadas se ele tivesse continuado o seu seguimento.

Neste trabalho considera-se o mecanismo de omissão completamente aleatória visto que se dispõe de dados não balanceados e considera-se que a falta de valores não parece estar associada à progressão da doença. Little e Rubin (2000) consideram que os dados

omissos provenientes deste tipo de mecanismo podem ser ignorados quando se adota a função de verosimilhança para a inferência dos modelos.

Para obter informação mais detalhada sobre dados omissos e os mecanismos referidos sugere-se a leitura de Diggle et al. (2002) e Little e Rubin (2000).

3.1.2 Variograma

O variograma de um processo estocástico $Y(t)$ é dado por (Diggle et al., 2002),

$$\gamma(u) = \frac{1}{2}\, Var\{Y(t) \; - \; Y(t-u)\}, \quad u \geq 0$$

Se $Y(t)$ é um processo estocástico, a função de autocorrelação $\rho(u)$ e a variância de $Y(t)$, σ^2, apresentam a seguinte relação,

$$\gamma(u) = \sigma^2\{1 - \rho(u)\}$$

O variograma empírico é estimado calculando-se a metade das diferenças observadas entre o par de resíduos, $v_{ij} = \frac{1}{2}\,(r_{ij} \; - \; r_{ik})^{\,2}$, e as diferenças do tempo correspondentes, $u_{ijk} = t_{ij} - t_{ik}$, onde $r_{ij} = \; Y_{ij} \; - \; \mu_{ij}$ e $j < k = 1, \ldots, m_i$. A função de autocorrelação em qualquer atraso u pode ser estimada, da seguinte forma, a partir do variograma amostral:

$$\hat{\rho}(u) = 1 - \frac{\hat{\gamma}(u)}{\hat{\sigma}^2}$$

sendo $\hat{\gamma}(u)$ a média de todos os v_{ij} associados aos valores particulares de u e $\hat{\sigma}^2$ é a variância estimada do processo.

3.1.3 Análise de diagnóstico

A análise de diagnóstico consiste na validação da estrutura de correlação escolhida, da homogeneidade da variância e da normalidade dos erros de medição. Para a validação da estrutura de correlação sobrepõem-se o variograma ajustado ao variograma empírico. Esta sobreposição permite avaliar, através da comparação gráfica, qual estrutura de correlação é mais adequada. O pressuposto da homogeneidade da variância dos erros de medição e o pressuposto da normalidade dos mesmos podem ser validados analisando-se a representação gráfica dos resíduos específicos do sujeito *versus* as respostas ajustada e a representação gráfica de um $Q - Q$ normal dos resíduos especificos do sujeito, respetivamente.

3.2 Análise de sobrevivência

Seja T uma variável aleatória não negativa que representa o tempo de sobrevivência de um indivíduo numa determinada população, ou seja, o tempo decorrido até ao evento de interesse. A probabilidade de um indivíduo sobreviver para além de um tempo t é dada pela função de sobrevivência,

$$S(t) = P(T > t),\ t \geq 0.$$

No caso contínuo, a função de risco (*hazard function*) que descreve a probabilidade instantânea de morte do indivíduo ao longo do tempo é,

$$\lambda(t) = \lim_{\triangle t \to 0^+} \frac{P(t \leq T < t + \triangle t | T > t)}{\triangle t}$$

Na análise de sobrevivência é bastante comum existir perda de informação temporal uma vez que raramente se acompanha todos os indivíduos em estudo até à ocorrência do evento de interesse e, por vezes, até o tempo de início é desconhecido. Esta perda de informação pode ser censura ou truncatura. A censura é uma característica muito frequente nos dados utilizados na análise de sobrevivência, o que a distingue dos outros métodos estatísticos. Este mecanismo ocorre quando apenas se tem conhecimento de que o evento de interesse ocorreu num dado intervalo de tempo. Quando o tempo até ocorrer o evento de interesse é maior que o tempo de observação, ou seja, quando apenas se sabe que o tempo excede determinado valor, está-se perante censura pela direita. Por outro lado, quando somente se sabe que o tempo é inferior a determinado valor está-se perante censura à esquerda. A censura intervalar ocorre quando apenas se sabe que o evento ocorreu num determinado período de tempo, ou seja, não se observou o momento exato da ocorrência do evento, mas sabe-se que não antecedeu um determinado valor nem excedeu um outro. Numa amostra com dados censurados, considera-se o tempo de censura potencial, c_i, e o indicador δ_i que toma o valor 1 se o indivíduo i observou o evento de interesse e 0 caso contrário. Assim, as observações são da forma (t_i, δ_i), $i = 1, \ldots, n$ onde $t_i = min(T_i, c_i)$ sendo T_i a v.a que representa o tempo de vida do i-ésimo indivíduo. A truncatura ocorre quando a perda de informação está relacionada com a exclusão de indivíduos do estudo por algum motivo relacionado com a ocorrência do evento. Considera-se truncatura à esquerda quando a perda de informação advém de excluir do estudo os indivíduos que já tinham observado o evento de interesse antes do início do estudo e não podiam ser observados. A truncatura à direita surge quando o critério de seleção dos indivíduos inclui somente os que sofreram o evento.

3.2.1 Estimação não paramétrica da função de sobrevivência

Se numa amostra de dimensão n não existirem observações censuradas, o estimador mais utilizado para a estimação de $S(t)$ é o estimador empírico,

$$\hat{S}(t) = \frac{número\ de\ observações > t}{n},\quad t \geq 0.$$

Kaplan e Meier (1958) apresentaram o estimador limite-produto, que estima a função de sobrevivência, tendo em conta as observações censuradas. $\hat{S}^{KM}(t)$ é uma função em escada com saltos nos instantes de ocorrência do evento de interesse. Este estimador não paramétrico é uma generalização do estimador empírico para dados censurado dado por,

$$\hat{S}^{KM}(t) = \prod_{t_i \leq t}(1 - \frac{d_i}{n_i})$$

onde d_i é o número de eventos ocorridos no tempo t_i e n_i o número de indivíduos em risco no tempo t_i.

3.2.2 Comparação de curvas de sobrevivência

Considerando-se duas amostras com m e n indivíduos e respetivas funções de sobrevivência $S_1(t)$ e $S_2(t)$, e pretende-se testar as seguintes hipóteses:

$$H_0 : S_1(t) = S_2(t) \text{ vs } H_1 : S_1(t) \neq S_2(t)$$

Quando se pretende comparar curvas de sobrevivência, a fim de se concluir sobre a sua igualdade, o teste mais utilizado é o teste de Log-Rank. Este é o teste mais potente, no entanto, a sua implementação só é recomendada quando o pressuposto de riscos proporcionais é verificado, uma vez que se isto não se verificar, este teste pode não permitir detetar diferenças significativas entre as curvas de sobrevivência. Uma maneira informal de se avaliar o pressuposto é analisando-se a representação gráfica das curvas de sobrevivência e certificando-se que estas não se cruzam. Se houver cruzamento das curvas o teste de Log-Rank não deve ser considerado (Rocha e Papoila, 2009). Com base no *software* *R*, não sendo possível aplicar-se o teste de Log-Rank, é implementado o teste de Gehan-Wilcoxon com a modificação de Peto e Peto. Estes dois testes fazem parte da classe de testes não paramétricos e apresentam a seguinte estatística de teste,

$$\frac{[\sum_{j=1}^{r} w_j(d_{1j} - e_{1j})]^2}{\sum_{j=1}^{r} w_j^2 v_{1j}}$$

onde w_j são constantes conhecidas que representam o peso atribuído a cada observação e dependem do teste utilizado. No teste de Log-Rank, $w_j = 1$ e no teste de Peto e Peto (1972), $w_j = \prod_{i:t(i)\leq t(j)} \left(1 - \frac{d_i}{n_i+1}\right)$. O número de eventos ocorridos no tempo t_j é dado por d_j e o número de indivíduos em risco no tempo t_j é dado por n_j. O valor médio e a variância condicionais a d_{1j} são, respetivamente:

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

Sob H_0, esta estatística tem uma distribuição assintótica de Qui-quadrado com um grau de liberdade.

Os testes apresentados são extensível a três ou mais grupos, permitindo a compação de r curvas de sobrevivência. Sob a hipótese nula $S_1(t) = \cdots = S_r(t)$, a estatística de teste obtida tem distribuição de Qui-quadrado com $r - 1$ graus de liberdade. Para informação mais detalhada sugere-se Klein e Moeschberger (1997) e Collett (2003).

3.2.3 Modelo de riscos proporcionais

Pretende-se, muitas vezes, perceber se o tempo de sobrevivência pode ser influênciado por variáveis que caracterizam de certa forma os indivíduos. Assim, é importante considerá-las no modelo, com o intuito de o melhorar. Um dos modelos mais frequentemente usados para perceber a influência de covariáveis na sobrevivência é o modelo de regressão de Cox (1972). Este é um modelo semi-paramétrico dado por,

$$h(t;\boldsymbol{z}) = \ h_0(t)\exp\{\beta_1 z_1 + \ \ldots + \beta_p z_p \ \}$$

onde $h_0(t)$ é a função de risco subjacente, ou seja, a função de risco comum a todos os indivíduos quando todas as covariáveis estão nulas, ou seja, $\boldsymbol{z} = \mathbf{0}$. $\boldsymbol{\beta}$ é o vetor de coeficientes de regressão (desconhecidos) e $\boldsymbol{z}$ é o vetor de covariáveis.

A fórmula apresentada é o produto de dois fatores, onde $h_0(t)$ representa a parte não paramétrica do modelo e $\exp\{\boldsymbol{\beta}'\boldsymbol{z}\}$ representa a parte paramétrica do mesmo, por este motivo trata-se de um modelo semi-paramétrico. O modelo exposto é conhecido como modelo de riscos proporcionais, na medida em que a razão das funções de riscos para dois grupos com vetores de covariáveis fixos é constante no tempo. Considerando os vetores de covariáveis $\boldsymbol{z_1}$ e $\boldsymbol{z_2}$, verifica-se que o risco relativo para estes é constante no tempo,

$$HR = \tfrac{h(t;\boldsymbol{z_1})}{h(t;\boldsymbol{z_2})} = \exp\{\boldsymbol{\beta}'(\boldsymbol{z_1} - \boldsymbol{z_2})\}$$

3.2.4 Análise de diagnóstico

A análise de diagnóstico consiste na avaliação do ajuste global do modelo final e na verificação da hipótese de riscos proporcionais, após o ajuste de um modelo de Cox aos dados em análise.

Para a avaliação do ajuste global do modelo final são geralmente utilizados os resíduos de Cox e Snell (1968). Considerando-se T uma v.a contínua com função de distribuição $F(T) \sim U(0,1)$ e função de sobrevivência $S(T) \sim U(0,1)$, vem que $H(T)$ tem distribuição exponencial de valor médio 1, visto que $H(T) = -logS(T)$.

O resíduo para o i-ésimo indivíduo, $i = 1, \ldots, n$ é dado por,

$$r_i = \hat{H}(t_i) = \exp(\hat{\beta}'\boldsymbol{z}_i)\hat{H}_0(t_i)$$

considerando que $H(t;\boldsymbol{z}) = \int_0^t h_0(u)\exp(\beta'\boldsymbol{z})du \ = \ \exp(\beta'\boldsymbol{z})H_0(t)$ e que $\hat{\beta}$ e $\hat{H}_0(t_i)$ representam as estimativas de máxima verosimilhança parcial. Os resíduos devem comportar-se como uma amostra proveniente de uma distribuição exponencial de valor médio 1, visto que se o modelo ajustado for aceitável, as estimativas $\hat{H}(t_i)$ apresentaram propriedades semelhantes a $H(t_i)$. Para dados censurados os resíduos sofrem uma alteração (r_i'), sendo acrescentado o valor 1 às observações censuradas e mantendo-se o valor r_i para as observações não censuradas.

A adequabilidade do modelo pode ser verificada através da representação gráfica dos pontos $(r_i',\ \widetilde{H}(r_i'))$, onde $\widetilde{H}(r_i')$ é a estimativa de Nelson-Aalen da função de risco cumulativa dos resíduos. Se os pontos representados se aproximarem da representação de uma reta de declive um e ordenada zero, concluí-se que os resíduos são proveniente de uma população Exp(1) e o modelo é adequado.

Para avaliar a hipótese dos riscos proporcionais, após o ajuste do modelos final de Cox pode-se utilizar os resíduos de Schoenfeld (1982). O resíduo de Schoenfeld associado à covariável z_j, $j = 1, \ldots, p$ para o i-ésimo indivíduo, é dada por

$$r_{ij} = \delta_i\{z_{ij}\ -\ a_{ij}\}$$

onde δ_i indica se a observação é censura ou não e

$$a_{ij} = \frac{\sum_{l \in R_i} z_{jl} \exp\{\hat{\beta}' \boldsymbol{z}_l\}}{\sum_{l \in R_i} \exp\{\hat{\beta}' \boldsymbol{z}_l\}}$$

Se a representação gráfica dos resíduos de Schoenfeld *versus* os tempos-até-evento tiveram um aspeto de uma nuvem aleatória de pontos, centrada em zero, pode-se considerar que o modelo tem um ajuste adequado.

Grambsch e Therneau (1994) apresentaram uma versão destes resíduos que garantem ter maior eficácia a detetar afastamentos do modelo assumido. Conhecidos como resíduos de Schoenfeld padronizados ou ponderados r_{ij}^*, estes são caracterizados como sendo as componentes do vetor

$$\boldsymbol{r}_i^* = k \times \mathrm{var}(\hat{\boldsymbol{\beta}})\boldsymbol{r}_i$$

onde se considera o vetor dos resíduos de Schoenfeld associado ao i-ésimo indivíduo $r_i = (r_{1i}, r_{2i}, \ldots, r_{pi})'$, k o número de observações não censuradas entre os n indivíduos e $var(\hat{\boldsymbol{\beta}})$ a matriz de covariância dos estimadores dos parâmetros β_j do modelo de Cox ajustado.

3.3 Análise modelos conjuntos

No contexto de modelos conjuntos é fundamental distinguir claramente através da notação os dois processos considerados: o processo longitudinal $\mathbf{Y}$ e o processo tempo-até-evento $\mathbf{F}$. Neste contexto, o interesse está na possibilidade destes dois processos estarem associados (Sousa, 2011). É comum existirem dados omissos em ambos os processos, quando o evento de interesse está em falta o processo de falha é chamado de censura $\mathbf{C}$ e quando falta uma medida longitudinal considera-se a existência de dados omissos $\mathbf{D}$. Dado que se considera que o processo $\mathbf{C}$ é independente do tempo de evento e do processo longitudinal, este processo é assumido como não informativo. Os modelos conjuntos modelam a distribuição conjunta $[\mathbf{Y}, \mathbf{F}]$, para $\mathbf{Y}$ e $\mathbf{F}$. A inferência sobre os parâmetros do modelo é feita por meio da decomposição da probabilidade total, mais especificamente

através da decomposição da verosimilhança total. Como a distribuição conjunta para as variáveis aleatórias consideradas não é clara, faz-se uso da regra de Bayes para a factorização dessa distribuição. Dependendo da forma que se fatoriza a verosimilhança total obtêm-se diferentes estratégias de modelos que levam a interpretações distintas, e consequentemente, a abordagens diferentes de problemas individuais,

Modelos de mistura de padrões:	Modelos de seleção:
$[\mathbf{Y},\mathbf{F}] = [\mathbf{F}][\mathbf{Y}\|\mathbf{F}]$	$[\mathbf{Y},\mathbf{F}] = [\mathbf{Y}][\mathbf{F}\|\mathbf{Y}]$

A interpretação dos parâmetros envolvidos em cada uma das componentes do modelos é diferente, pois num modelo eles referencem-se à distribuição condicional e no outro à distribuição marginal. Tendo em conta a natureza do problema estatístico e as questões científicas a serem respondidas, adopta-se um tipo de modelo. A interpretação estatística dos modelos é diferente apesar de os modelos, matematicamente, descreverem exatemente a mesma distribuição. Quando o foco da inferência reside nos parâmetros do modelo de tempo-até-evento e se permite haver correlação nas medidas repetidas, opta-se pelos modelos de seleção. Quando o foco está na trajetória longitudinal mas se permite a associação a um padrão de evento, usam-se os modelos de mistura de padrões. Desta forma, o entendimento e a inferência dos parâmetros do modelo são diferentes para as duas abordagens. Estes modelos podem incorporar efeitos aleatórios, e são chamados de modelos de mistura de padrões aleatórios e modelos de seleção aleatória. Nos modelos de mistura de padrões os efeitos aleatórios são incorporados na distribuição marginal dos tempos de evento e nos modelos de seleção, os efeitos aleatórios são incorporados no modelo longitudinal marginal. Portanto, tendo em conta os efeitos aleatórios, as distribuições conjuntas tem a seguinte forma:

Modelos de mistura de padrões aleatórios:	Modelos de seleção aleatória:
$[\mathbf{Y},\mathbf{F},\mathbf{U}] = [\mathbf{U}][\mathbf{F}\|\mathbf{U}][\mathbf{Y}\|\mathbf{F}]$	$[\mathbf{Y},\mathbf{F},\mathbf{U}] = [\mathbf{U}][\mathbf{Y}\|\mathbf{U}][\mathbf{F}\|\mathbf{Y}]$

Para além dos modelos de mistura de padrões aleatórios e dos modelos de seleção aleatória existem também os modelos conjuntos de efeitos aleatórios definidos por Diggle (1998). Estes são uma classe diferente de modelos conjuntos que assumem que as medições repetidas e o tempo-até-evento dependem de um efeito aleatório não observado, especificado através duma distribuição bivariada. Dados os efeitos aleatórios $\mathbf{U} = (\boldsymbol{U_1}, \boldsymbol{U_2})$, assume-se independência condicional entre $\mathbf{Y}$ e $\mathbf{F}$, e os modelos conjuntos de efeitos aleatórios são descritos como,

$$[\mathbf{Y},\mathbf{F},\mathbf{U}] = [\mathbf{U}][\mathbf{Y}|\boldsymbol{U_1}][\mathbf{F}|\boldsymbol{U_2}]$$

A estrutura de correlação entre U_1 e U_2 determina a associação existente entre os dois processos neste tipo de modelo.

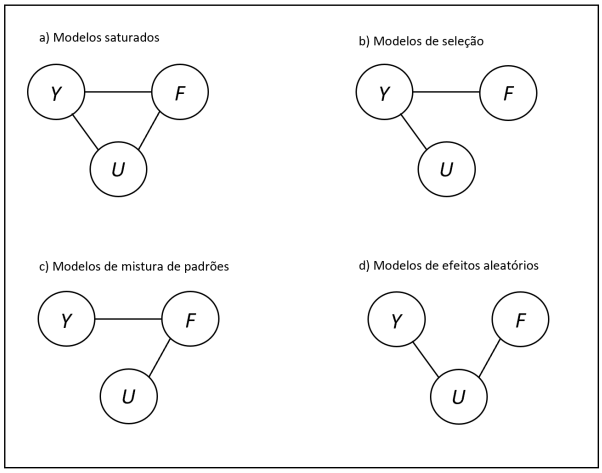

Figura 3.1: Representação gráfica de modelos saturado, modelos de seleção, modelos de mistura de padrões e modelos de efeitos aleatórios

Tal como em Sousa (2011) os diagramas na Figura 3.1 representam a independência condicional para as três variáveis aleatórias referidas anteriormente. A ausência de uma linha indica independência condicional entre os dois processos, dado o terceiro processo considerado. A Figura 3.1 (a) representa o modelo saturado, onde todas as associações são possíveis. A Figura 3.1 (b) representa os modelos de seleção, onde as medidas longitudinais são influenciadas pelos seus efeitos aleatórios individuais, mas estes não influênciam o processo **F**. Por outro lado, nos modelos de mistura de padrões da Figura 3.1 (c) os efeitos aleatórios individuais determinam o tempo do evento, que após ser predefinido desenvolve o perfil longitudinal individual. Nos modelos conjuntos de efeitos aleatórios, Figura 3.1 (d) os processos são considerados uma resposta conjunta a um processo específico individual não observado condicionado ao facto de haver independência entre as respostas.

Neste trabalho será utilizada a abordagem de modelos de efeitos aleatórios, uma vez que o principal interesse da presente análise reside na associação entre o processo longitudinal e o processo de sobrevivência. Para uma explicação mais detalhada sugere-se a leitura de Sousa (2011) de onde foi recolhida toda a informação exposta nesta secção.

O facto da resposta longitudinal e dos mecanismos ausentes serem modelados partilhando efeitos aleatórios faz com que os modelos de efeitos aleatórios sejam também conhecidos como modelos de parâmetros partilhados. Em modelos conjuntos de efeitos aleatórios, assume-se que quer o tempo do evento como o processo longitudinal dependem de um efeito aleatório, que pode ser entendido como um processo não observável, como uma doença subjacente que influência ambos os processos. Para além disso, quando condicionados aos efeitos aleatórios, os dois processos são independentes. Segundo Hogan e Laird (1997) a distribuição conjunta pode ser definida como,

$$[\boldsymbol{Y},\boldsymbol{F}]=\int_{\boldsymbol{U}} [\boldsymbol{U}][\boldsymbol{Y}|\boldsymbol{U}][\boldsymbol{F}|\boldsymbol{U}]d\boldsymbol{U}$$

Habitualmente utilizam-se modelos lineares de efeitos mistos, com efeitos aleatórios ao

nível individual para a modelação das medidas repetidas, e modelos de regressão de Cox com fragilidade log-Gaussiana para a parte do processo de tempo de falha. Ao se permitir que os efeitos aleatórios gaussianos do modelo linear estejam correlacionados com o termo de fragilidade do modelo de regressão de Cox modela-se a dependência estocástica.

A variável latente U, pode ser definida como sugerido por Henderson, Diggle, e Dobson (2000), da seguinte forma,

$$U_i(t) = U_{0i} + U_{i1}t_{ij}$$

onde U_{0i} e U_{1i} representam, para o sujeito i, os efeitos aleatórios de interceção e inclinação, respetivamente.

3.3.1 *Packages* utilizados na análise conjunta disponíveis no *software R*

Foram utilizados dois *packages* disponíveis no *software R* que adotam a metodologia de efeitos aleatórios desenvolvida por Wulfsohn e Tsiatis (1997) com a extensão de Henderson et al. (2000). Dependendo do foco da análise a ser desenvolvida, escolhe-se o *package* que melhor responderá às questões que se quer responder (McCrink, Marshall, e Cairns, 2013). O modelo conjunto implementado no *package JM* desenvolvido por Rizopoulos (2010), foca-se mais no processo de sobrevivência e pretende perceber como este é influenciado por uma covariável longitudinal. O *package joineR* desenvolvido por Philipson et al. (2012), permite inferir sobre a força de ligação que existem entre os dois processos quando se implementa um modelo conjunto. Neste trabalho utilizaram-se ambos os *packages*, pois para além de ambos os objetivos serem de interesse para a presente análise, pretende-se perceber as diferenças existentes nos resultados obtidos nos diferentes *packages*.

Para o processo longitudinal é implementado o mesmo modelo linear de efeitos aleatórios, em ambos os *packages*:

$$Y_{ij} = \mu_{ij} + U_i + Z_{ij} = X_1\beta + U_{0i} + U_{1i}t_{ij} + Z_{ij}$$

onde X_1 é a matriz de desenho para as covariáveis fixas, com os parâmetros de regressão β correspondentes. Na presente análise partiu-se de um modelo longitudinal saturado e foi-se retirando as variávies que não se mostravam significativas no processo longitudinal da abordagem de modelos conjuntos. A variável latente dada por (U_{0i},U_{1i}) é uma realização de uma MVN(0,Σ) onde $\Sigma = \begin{pmatrix} v_1^2 & v_{12} \\ v_{12} & v_2^2 \end{pmatrix}$ é a matrix variância/covariância dos efeitos aleatórios. Z_{ij} são realizações independentes e identicamente distribuídas de uma N(0,τ^2), que representa a variabilidade não explicada.

Os dois *packages* diferem, essencialmente, na forma como é incorporada a variável aleatória latente no processo de sobrevivência.

No *package JM* a estimativa do processo longitudinal não observado, $m_i(t) = \hat{\mu_{ij}} + \hat{U_i}$, é incorporada no modelo de sobrevivência da seguinte forma,

$$h_i(t) = h_0(t) \exp\{X_{2i}\beta_2 + \ \alpha m_i(t)\}$$

onde X_{2i} representa as covariáveis em *baseline*, neste estudo, as covariáveis com efeito significativo na análise de sobrevivência quando se analisou apenas o processo de sobrevivência, e β_2 o vetor dos parâmetros de regressão correspondentes. O efeito que a resposta longitudinal tem no processo de sobrevivência é determinado pelo parâmetro α. Neste estudo as respostas longitudinais são o valor do marcador tumoral CA15-3 e, também, o valor do marcador tumoral CEA, ambos na escala logarítmica.

No *package joineR* os efeitos aleatórios longitudinais são incorporados no modelo de sobrevivência da seguinte forma:

$$h_i(t) = h_0(t) \exp\{X_{2i}\beta_2 + \gamma_0 U_{0i} \ + \gamma_1 U_{1i} t\}$$

onde γ_0 e γ_1 representam os parâmetros de associação, sendo respetivamente, o efeito da interceção aleatória longitudinal e a inclinação no processo de sobrevivência.

Como a incorporação dos efeitos aleatórios é diferentes nos dois *packages*, pode acontecer que os valores estimados de β_2 sejam diferentes e que as próprias covariáveis tenham um efeito estatísticamente significativo na sobrevivência do sujeito dentro do modelo conjunto ajustado com o *package joineR*, mas nenhum efeito significativo no modelo conjunto ajustado com o *package JM* ou vice-versa. Enquanto que com o *package JM* consegue-se perceber que fatores influenciam a mudança nos valores dos marcadores tumorais considerados nesta análise e de que forma influenciam a sobrevivência do paciente, com o *package joineR* consegue-se determinar o efeito da especificidade do indivíduo na sua sobrevivência, γ_0, e de que forma a expecificidade do sujeito altera a sobrevivência ao longo do tempo, γ_1.

É importante salientar que o *software* usado apresenta duas limitações. Apesar do modelo de efeito aleatório comum permitir lidar com dados de censura à direita e truncados à esquerda, os dois *packages* usados ainda não são capazes de lidar com dados truncados à esquerda. Portanto, considera-se apenas um modelo de Cox que incorpora a censura à direita no processo de sobrevivência. Desta forma, é relevante apenas comparar os resultados das estimativas do processo de sobrevivência com os resultados do modelo que considera este mesmo mecanismo de censura na análise efetuada em separado. A outra restrição diz respeito ao fato de não ser possível escolher a estrutura de correlação que melhor descreve a variabiliade intra-indivíduos, como acontece na análise longitudinal efetuada em separado.

3.3.2 Análise de diagnóstico

Para o diagnóstico sobre o ajuste do modelo, representa-se gráficamente os resíduos específicos do sujeito *versus* as respostas ajustadas. Este gráfico é usado para avaliar o pressuposto de variância constante de Z_{ij}. Além disso, um gráfico $Q - Q$ normal dos resíduos específicos do sujeito é usado para verificar o pressuposto de normalidade de Z_{ij}. Para validar o processo de sobrevivência ajustada, utiliza-se a representação gráfica dos resíduos de Cox e Snell (1968), já explicada em detalhe anteriormente.

Capítulo 4

Análise estatística dos dados do cancro da mama

Neste capítulo apresentam-se os resultados obtidos na aplicação dos modelos, apresentados no capítulo anterior, aos dados do cancro da mama. Começa-se por apresentar a análise exploratória dos dados, onde são mencionados os principais riscos para desenvolvimento do cancro da mama e de que forma é caractereizada a base de dados utilizada tendo em conta esses riscos. Na análise longitudinal, são apresentados os resultados das análises efetuadas em separado para os dois marcadores considerados: CEA e CA15-3. Na análise de sobrevivência, são estimadas curvas de Kaplan-Meier e apresenta-se um modelo de regressão de Cox. Por fim, são apresentados os resultados das análises conjuntas para os dois marcadores tumorais em estudo.

4.1 Análise exploratória

A base de dados inicial consistia na informação recolhida de 596 pacientes, com diagnóstico de cancro da mama entre os anos de 2008 e 2012, no Hospital de Braga, e de todos os pacientes em acompanhamento à data de 1 de janeiro de 2008. Destes pacientes, 56 foram excluídos por apresentarem, pelo menos, um dos seguintes critérios de exclusão:

- Género masculino
- Neoplasia benigna da mama
- Falta de informação de diagnóstico, tratamento ou seguimento

Após análise clínica individual, o médico assistente indicou que os 19 casos de cancro bilateral deveriam ser analisados de forma independente. O facto de existirem pacientes com cancro bilateral que observaram recidiva da doença em apenas uma das mamas é uma razão para se assumir esta independência. Desta forma, há 559 pacientes do género

feminino diagnosticadas com um tumor maligno da mama que serão consideradas para as análises futuras sendo estes a soma dos 540 pacientes com os 19 casos referidos de cancro bilateral.

A Unidade de Senologia do Hospital de Braga foi criada em 2008. No entanto, pacientes com data de diagnóstico anterior a 2008 que estavam em acompanhamento no hospital foram considerados nesta amostra. Há diagnósticos desde 1993. Esta informação é importante uma vez que, por exemplo, na análise de sobrevivência deve-se considerar truncatura à esquerda, visto que apenas está disponível informação das pacientes que se encontram vivas, não havendo registo das pacientes diagnosticadas antes de 2008 e que faleceram. Assim, pode-se considerar que a base de dados está dividida em dois grupos: o grupo de pacientes que foram diagnosticadas antes de 1 janeiro de 2008 e um outro grupo de pacientes que foram diagnosticadas depois desta data. Das 186 pacientes que foram diagnosticadas antes de 1 de janeiro de 2008, 46 tiveram recidiva da doença enquanto 37 pacientes das 373 diagnosticadas após essa data observaram a recidiva do cancro da mama. Há 83 casos de recidivas sendo 11 recidivas locais, 64 recidivas metastizadas e 6 recidivas locais e metastizadas. Não há informação sobre o tipo de recidiva de duas pacientes (Anexo B). Assim, não se observou recidiva em 478 pacientes. No entanto, como o médico consultor indica, geralmente no momento anterior à morte há associada uma recidiva, isto faz com que a classificação "sem recidiva"seja ambígua. Considerando-se as 559 pacientes, há um total de 60 mortes onde 52 destas estão diretamente relacionadas com a recidiva da doença.

Tabela 4.1: Variáveis recolhidas na Unidade de Senologia do Hospital de Braga

Variáveis exploratórias ao nível individual	**Variáveis exploratórias ao nível do tumor**
Data de nascimento	Tipo histórico de tumor
Estado civil	Grau de diferenciação- Bloom-Richardson
Distrito de residência	Presença de carcinoma associado
País de residência	Imagens de invasão linfática
Freguesia de residência	Imagens de invasão venosa
Profissão	Recetores de estrogénio
Habilitações literárias	Recetores de progesterona
Idade da menarca	HER-2/neu
Número de filhos	Ki-67 índice proliferativo
Idade na primeira gravidez	Tratamento neoadjuvante
Amamentação	Tratamento cirúrgico
Duração da amamentação (meses)	Biopsia de linfa sentinela
Menopausa	Resultados da biopsia de linfa sentinela
Tipo de menopausa	Dissecção axilar
Idade da menopausa	Resultado da dissecção axilar
Histórico familiar de cancro da mama	Tratamento pós-cirúrgico
Grau de parentesco	Mama afetada
Substituição hormonal terapêutica	Quadrante da mama
Contracetivos hormonais orais	Estado do tumor (TNM)
Idade de diagnóstico	Tamanho do primeiro tumor
	Grau de disseminação para nódulos linfáticos regionais
	Presença de metastases distantes
	Recidiva
	Tipo de Recidiva
	Valores de CEA
	Datas de recolha dos valores CEA
	Valores de CA 15.3
	Datas de recolha dos valores CA 15.3
	Datas de consultas de acompanhamento ou óbito
	Estado vital

Estão disponíveis mais de 50 variáveis (Tabela 4.1) que podem ser divididas em dois grupos:

- **Variáveis exploratórias ao nível individual** como característica demográficas e fatores de risco para desenvolvimento do cancro da mama apresentados por Rodrigues (2011) com base no trabalho de Trichopoulos et al. (2008), como por exemplo: cancro bilateral, histórico familiar de cancro, idade da menarca, tipo de menopausa, etc. Estas variáveis foram apenas medidas em *baseline*, ou seja, no momento do diagnóstico.

- **Variáveis exploratórias ao nível do tumor** que incluem características relacionadas com tumor e vários fatores prognósticos resumidos e reportados na literatura por Fitzgibbons et al. (2000) e Cianfrocca e Goldstein (2004) tais como subtipo HER2/neu, recetores hormonais (estrogénio e progesterona), entre outros. Estão disponíveis variáveis medidas em *baseline*, variáveis correspondentes ao evento de interesse (recidiva do cancro da mama) e variáveis longitudinais.

Nos anexos A, B, C e D são apresentadas as principais estatísticas das variáveis ao nível individual e ao nível do tumor cuja análise será apresentada em seguida.

Relativamente ao distrito de residência das pacientes, 96,42% das pacientes pertencem ao distrito de Braga, como era esperado tendo em conta a localização do hospital. Ainda assim, há 16 pacientes que pertencem ao distrito de Viana do Castelo (2,86%) e 4 ao distrito do Porto (0,72%).

Quase metade das pacientes tem um nível educacional igual ou inferior ao $4^{\underline{o}}$ ano (49,37%), havendo apenas 42 pacientes com escolaridade ao nível superior (7,51%). Não há informação sobre a escolaridade de 23,08% das pacientes.

Em seguida, descrevem-se sucintamente alguns riscos para desenvolvimento do cancro da mama mencionados por Rodrigues (2011). O objetivo futuro é identificar quais destes riscos estão relacionados com a recidiva do cancro da mama.

Pacientes diagnosticadas com cancro da mama antes dos 35 anos de idade são referidas como pacientes com prognóstico desfavorável, mesmo tendo em conta outros fatores. Desta forma a idade ao diagnóstico deve ser usada como complemento a outros fatores prognósticos com o objetivo de identificar um grupo de pacientes com maior risco de recidiva (Cianfrocca e Goldstein, 2004). A idade ao diagnóstico está compreendida entre os 20 e os 92 anos, sendo a mediana de 58 anos. Cerca de 3% das pacientes foram diagnosticadas antes dos 35 anos. A Figura 4.1 apresenta a distribuição da idade ao diagnóstico.

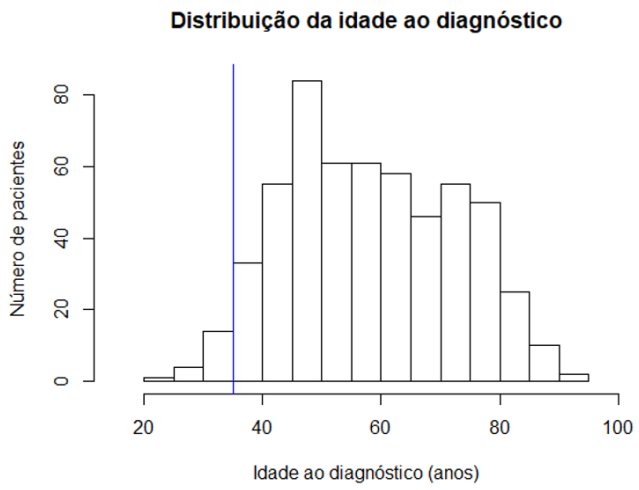

Figura 4.1: Histograma da idade ao diagnóstico

Mulheres com menarca precoce apresentam um maior risco de desenvolver cancro da mama, na amostra utilizada, 18,6% das pacientes estão nesta condição. Há 238 pacientes sem informação relativa a este fator etiológico. A idade da menarca compreende-se entre os 9 e os 19 anos, sendo a mediana de 13 anos. Cerca de 70% das pacientes tem idade de menarca entre os 11 e os 14 anos. A Figura 4.2 apresenta a distribuição da idade da menarca.

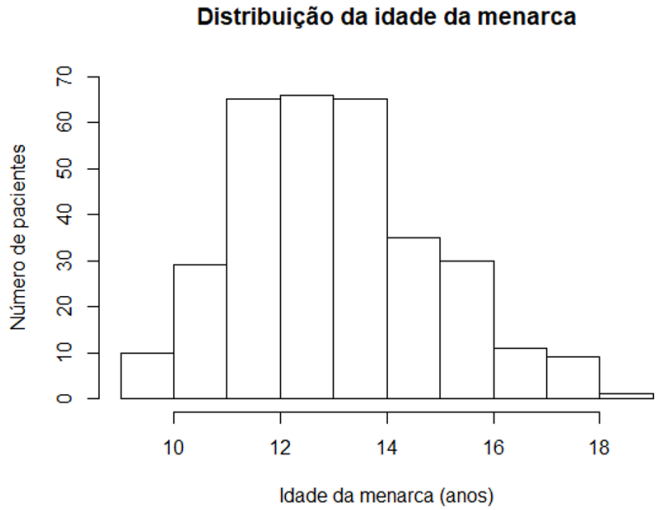

Figura 4.2: Histograma da idade da menarca

Relativamente à menopausa, 61,36% das pacientes estão em menopausa não havendo informação para apenas 3,76% das pacientes. A indução da artificial da menopausa parece estar associada a uma proteção em comparação com a menopausa natural, 47,41% das menopausas desta amostra são menopausas naturais. A Figura 4.3 apresenta a dis-

tribuição da idade da menopausa estando esta compreendida entre os 33 e os 58 sendo a mediana de 50 anos. Idade avançada da menopausa está associada a um maior risco de desenvolver cancro da mama, cerca de 4,11% das pacientes entraram na menopausa depois dos 55 anos.

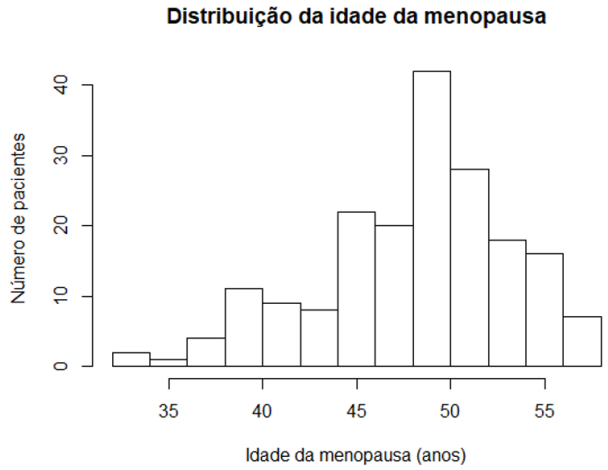

Figura 4.3: Histograma da idade da menopausa

Outro risco etiológico mencionado é a baixa paridade. Analisando-se a Figura 4.4, observa-se que 41 pacientes (7,33%) não tem filhos, 132 pacientes (23,61%) tem dois filhos e apenas 42 pacientes (7,51%) tem mais de 4 filhos. Não há informação para 36,67% das pacientes.

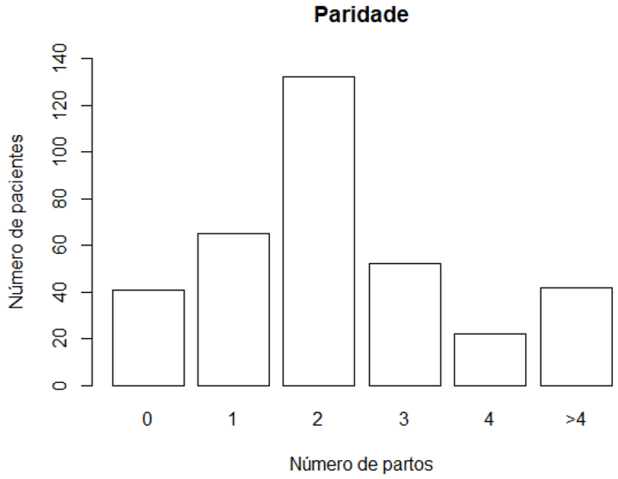

Figura 4.4: Gráfico de barras da paridade

A experiência reprodutiva em idades mais jovens induz uma diminuição do risco. A distribuição das idades à primeira gravidez é apresentada na Figura 4.5, salienta-se, no entanto, que não há informação 59,92% das pacientes. A idade varia entre os 15 e os 40

anos, sendo a mediana de 25 anos. Apenas 10 pacientes tiveram o primeiro filho após os 35 anos representando 1,79% da amostra.

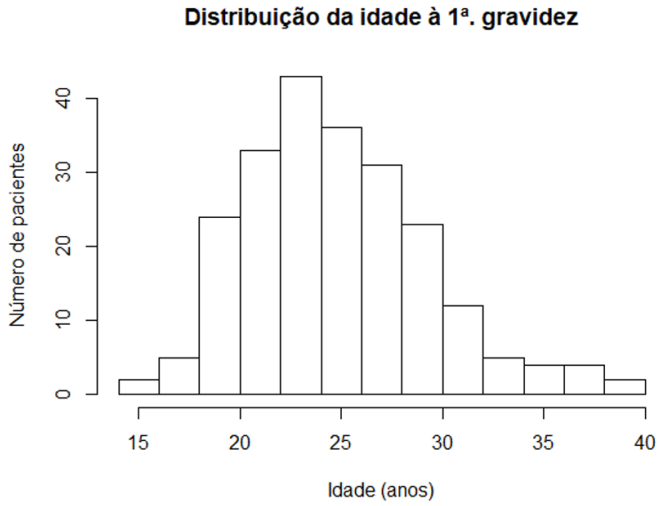

Figura 4.5: Histograma da idade da primeira gravidez

Relativamente à amamentação, não há informação para 336 pacientes (60,11%), ainda assim, 209 pacientes (37,39%) amamentaram e apenas 14 pacientes (2,5%) não amamentaram.

A terapêutica hormonal de substituição e o uso de contraceção hormonal oral são segundo Rodrigues (2011) fatores de risco para desenvolver cancro da mama. As percentagens de valores omissos para estas duas variáveis são muito elevados, verificando-se 83,01% e 67,08%, respetivamente. Apesar disso, 4,65% das pacientes foram submetidas a substituição hormonal terapêutica e 20,39% usaram contraceção hormonal oral.

A existência de histórico familiar de cancro é um fator etiológico. Estima-se que o risco aumenta cerca de 80% em pacientes com cancro em familiares de primeiro grau (Rodrigues, 2011). Não obstante a elevadíssima percentagem de dados omissos da variável (88,01%), observa-se na Figura 4.6 que 18 pacientes (3,22%) têm histórico familiar de primeiro grau.

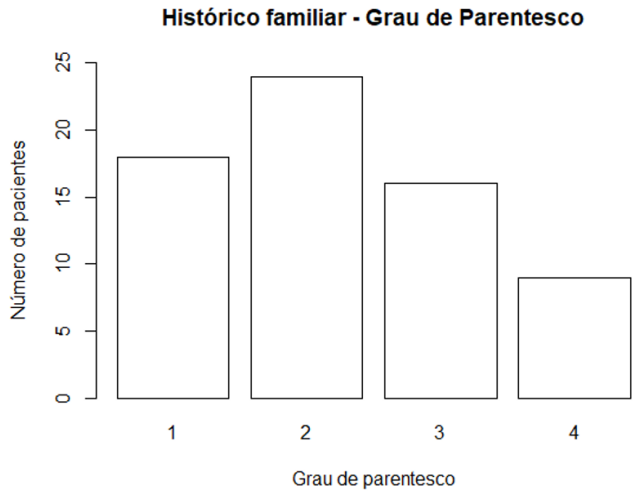

Figura 4.6: Gráfico de barras para o grau de parentesco

Como já havia sido referido, há 38 casos de cancro bilateral (ou seja, 19 pacientes que apresentam tumor maligno em ambas as mamas, não necessariamente ao mesmo tempo). Salienta-se que estas pacientes serão analisadas de forma independente como sugerido pelo médico assistente.

O tratamento do cancro da mama é definido pelo médico assistente (ou por uma equipa de médicos) e este tem várias opções disponíveis sendo o tratamento mais comum a cirurgia. Segundo o documento *Recomendações Nacionais para Diagnóstico e Tratamento do Cancro da Mama* dada a evolução associada ao tratamento deste tipo de cancro, as abordagens tendem a ser cada vez menos invasivas. Como técnica alternativa à mastectomia, na maioria das doentes com com carcinoma ductal in situ e carcinoma invasor em estadios iniciais da doença aceita-se a cirurgia conservadora da mama. A mastectomia deve ser considerada quando existem fatores que aumentam o risco de recidiva local, existem comorbilidades ou contra-indicações absolutas que impeçam a realização de radioterapia ou quando é preferência da doente. Nesta amostra do total das pacientes, 545 (97,5%) foram sujeitas a intervenção cirúrgica. Destas 224 fizeram mastectomia conservadora da mama e 321 fizeram mastectomia com ou sem dissecção dos linfonodos axilares. Das 14 pacientes que não foram submetidas a cirurgia apenas 3 foram submetidas a tratamento primário havendo um total de 11 pacientes que não tiveram nem tratamento primário nem cirurgia. Salienta-se que as 6 pacientes que apresentam recidiva local e metastizada foram sujeitas a mastectomia com ou sem dissecção dos linfonodos axilares. Como tratamentos neoadjuvantes, ou seja, pós- cirúrgicos, existe a quimioterapia, a radioterapia e a terapia hormonal. Nesta amostra, 66,01% das pacientes foram submetidas a quimioterapia, 66,19% foram submetidas a radioterapia e 85,51% foram submetidas a terapia hormonal. A percentagem de valores omissos são de 1,61%, 2,5% e 3,22%, respetivamente. Quase metade dos casos (44,72%) foram submetidas aos três tratamentos. Das pacientes que não

foram intervencionadas cirurgicamente nem tiveram tratamento primário apenas uma teve os três tratamentos neoadjuvantes e apenas uma não teve qualquer tipo de tratamento.

O sistema de estadiamento utilizado no cancro de mama foi proposto pelo *American Joint Committee on Cancer*, conforme explicado por de Oliveira e da Silva (2011). Nesse sistema, T define a extensão do tamanho do tumor, N define o estado nodal e M indica a presença ou ausência de metástases à distância. A classificação da doença oncológica em estadios - de 0 a IV - resulta do agrupamento das categorias TNM. Esta classificação é muito relevante dado que os fatores patológicos refletidos nesta classificação são muito importantes para ajudarem os médicos a estabelecerem um prognóstico e definirem um tratamento adequado. A Tabela 4.2 apresenta a explicação de cada categoria das diferentes variáveis referidas, sendo uma adaptação da tabela apresentada em de Oliveira e da Silva (2011). Adoptou-se a adaptação apresentada uma vez que esta já havia sido usada no estudo anteriormente desenvolvido com esta base de dados onde se estudou a morte como evento de interesse em Borges (2015).

Tabela 4.2: Sistema de estadiamento (TNM)

Sistema de estadiamento (TNM) para o cancro de mama do *American Joint Committee on Cancer*	
Tumor primário (T)	
TX	Tumor primário não encontrado
T0	Não há evidência de tumor primário
T1	Tumor não excede 2.0 cm na sua maior dimensão
T2	Tumor maior que 2.0 cm mas menor que 5.0 cm na sua maior dimensão
T3	Tumor maior que 5.0 cm na sua maior dimensão
T4	Tumor de qualquer tamanho com extensão direta para a parede torácica ou pele
Tis	Carcinoma intraductal, carcinoma lobular in situ, doença de Pagetś do mamilo sem invasão tumoral no tecido mamário normal.
Grau de disseminação para nódulos linfáticos regionais (N)	
NX	Os nódulos linfáticos regionais não podem ser avaliados (não removidos para estudo ou removidos anteriormente)
N0	Sem metástases nos nódulos linfáticos regionais
N1	Metástase em 1-3 nodulos linfáticos axilares ipsilaterais e / ou em nódulos linfáticos mamários internos ipsilaterais com metástase microscópica detetada por dissecção de nódulos linfáticos sentinela, mas não clinicamente aparente.
N2	Metástase em 4-9 nódulos linfáticos axilares ipsilaterais ou em nódulos linfáticos mamários internos ipsilaterais clinicamente aparentes na ausência de metástases em nódulos linfáticos axilares.
N3	Metástase em 10 ou mais nódulos linfáticos axilares ipsilaterais; ou em nódulos linfáticos infraclaviculares ipsilaterais; ou em nódulos linfáticos mamários internos ipsilaterais clinicamente aparentes na presença de um ou mais nódulos linfáticos axilares positivos; ou em mais de 3 nódulos linfáticos axilares com metástases microscópicas clinicamente negativas em nódulos linfáticos mamários internos; ou em nódulos linfáticos supraclaviculares ipsilaterais.
Metástases distantes (M)	
M0	Sem metástases distantes
M1	Com metástases distantes
* Neste estudo fez-se uso da classificação patológica, comumente denominada "pN", uma classificação baseada na dissecção de linfonodo axilar com ou sem biópsia de linfonodo sentinela, que mantém o mesmo significado para as categorias.	

As categorias T, N e M serão em seguida analisadas em separado (Anexo B). Relativamente ao tamanho do tumor (T), 47,94% são classificados como T1 e 36,67% são classificados como T2. Isto significa que a maior parte dos casos tem tumores identificados com medidas não superiores a 5 cm na sua maior dimensão. Apenas 10 casos (1,79%) não tem informação disponível a cerca do tumor primário. Em relação à classificação dos

nódulos linfáticos (N) 48,12% das pacientes não apresentam metástases nos linfonodos regionais. Há 20 casos sem classificação o que representa 3,58% do total da amostra. No que diz respeito à presença de metástases (M) apenas 4 casos (0,72%) apresentam metástases à distância, havendo 11 casos (1,97%) sem identificação para esta categoria.

Tabela 4.3: Estadios do *American Joint Committee on Cancer*

Grupos		
Estadio 0		Tis, N0, M0
Estadio I	Estadio IA	T1,N0,M0
	Estadio IB	T0,N1mi,M0
		T1,N1mi,M0
Estadio II	Estadio IIA	T0,N1,M0
		T1,N1,M0
		T2,N0,M0
	Estadio IIB	T2,N1,M0
		T3,N0,M0
Estadio III	Estadio IIIA	T0,N2,M0
		T1,N2,M0
		T2,N2,M0
		T3,N1,M0
		T3,N2,M0
		T4,N0,M0
	Estadio IIIB	T4,N1,M0
		T4,N2,M0
	Estadio IIIC	Qualquer T,N3,M0
Estadio IV		Qualquer T, qualquer N,M1

Da combinação das categorias T, N e M resulta a classificação do tumor em estadios que é apresentada na Tabela 4.3. Há um total de 193 casos (34,53%) com Estadio I e 213 casos (38,10%) Estadio II. Vinte e quatro casos (4,29%) não apresenta qualquer Estadio associado.

De acordo com o Anexo C onde é apresentada a distribuição dos tipos de carcinomas em estudo, verifica-se que 359 são carcinomas ductal invasores representado 64,22% dos casos. Este tipo de carcinoma representa 63,64% das recidivas locais, 71,43% das recidivas metastizadas e 80% das recidivas locais e metastizadas. Salienta-se ainda que das 6 recidivas locais e metastizadas existentes 5 apresentam carcinoma invasor (4 apresentam o carcinoma ductal invasor, 1 apresenta o tumor invasor misto lobular e 1 não tem carcinoma identificado).

O estudo desenvolvido por Hwang et al. (2017), demonstrou que invasões vasculares e linfáticas positivas apresentam taxas mais altas de recidivas local, regional e distante, apesar dos impactos serem mais proeminente em recidivas distantes. Nesta amostra 17,53% das pacientes apresentam imagens de invasão vascular linfática e 4,65% apresentam imagens de invasão vascular venosa.

De facto nas pacientes que apresentam invasão vascular linfática, a percentagem de mulheres com recidiva é maior. Há uma maior percentagem de mulheres com recidivas metastizadas e locais e metastizadas, em simultâneo, apesar de haver menor percentagem de mulheres com recidivas locais. (Figura 4.7)

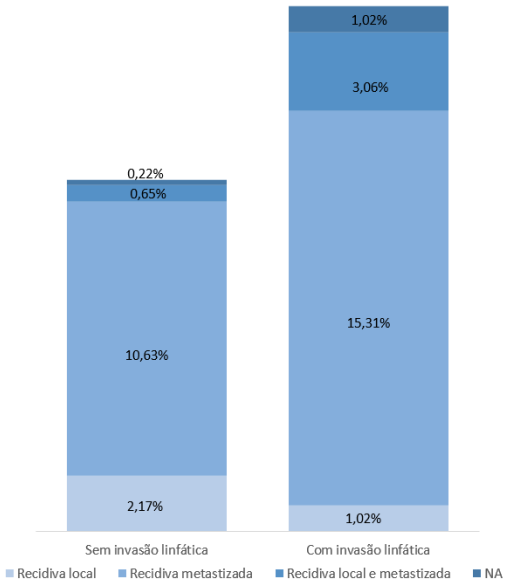

Figura 4.7: Invasão vascular linfática vs Tipo Recidiva

Das mulheres que apresentam invasão vascular venosa, a diferença de percentagem de mulheres sem recidiva é bastante mais evidente. Apesar de não haver nenhuma mulher com recidiva local, a percentagem de recidivas metastizadas e locais e metastizadas é bastante mais elevada nas mulheres com invasão vascular venosa em comparação com mulheres sem as invasões (Figura 4.8).

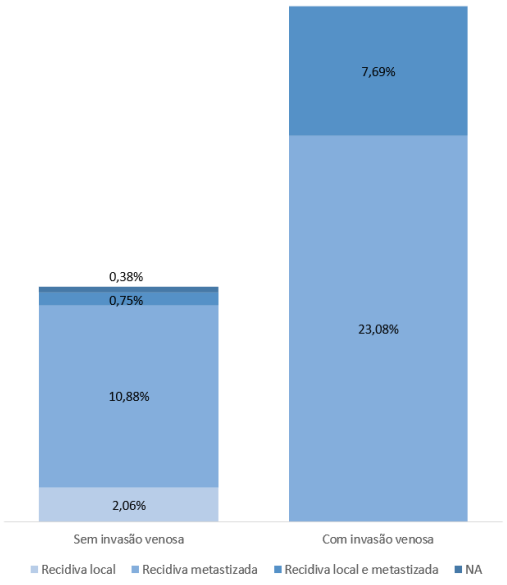

Figura 4.8: Invasão vascular venosa vs Tipo Recidiva

A classificação proposta por Bloom e Richardson (1957), apresenta três categorias: Grau I (baixa malignidade); Grau II (malignidade intermédia) e Grau III (alta malignidade). Esta classificação tem como objetivo refletir a malignidade potencial do tumor e indicar os casos que no momento do tratamento, têm maior predisposição para apresentar metástases ocultas à distância. Contudo, aquando do diagnóstico, as metástases parecem estar presentes nos três graus. Assim, na prática, esta graduação histológica fornece um guia para a velocidade com que essas metástases se tornam ativas, produzem sintomas e causam a morte. Na nossa amostra a classificação Gx significa que não foi possível identificar o grau do tumor. Conforme se verifica na Figura 4.9, 224 pacientes (40,07%) tem o seu tumor classificado como Grau II havendo um total de 47 casos (8,41%) sem informação disponível. Das mulheres que estavam classificadas com grau III, cerca de 19% apresentou recidiva metastizada da doença. Nos restantes graus a percentagem deste tipo de recidiva é bastante inferior.

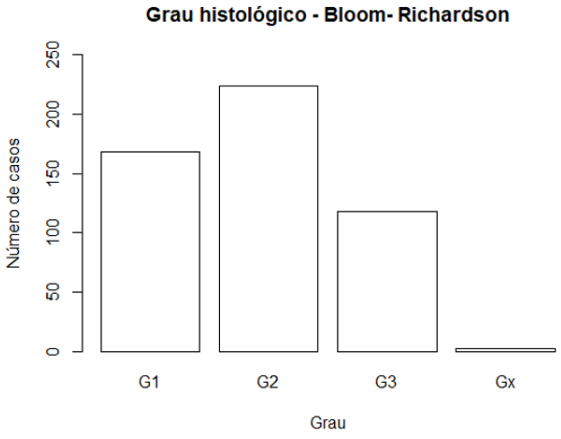

Figura 4.9: Gráfico de barras para o grau histológico de Bloom e Richardson

De acordo com os exames adicionais para diagnóstico do cancro da mama, mencionados no site da Liga Portuguesa Contra o Cancro[1], em todos os casos de cancro da mama são feitos testes aos recetores hormonais (estrogénio e progesterona) e deve ser analisado o aumento (ou sobre-expressão) do receptor-2 para o fator de crescimento epidérmico humano (HER2) que é o recetor que existe na membrana das células tumorais, também designado gene HER2/neu. O cancro da mama HER2+ é um subtipo específico do cancro e está associado a maior agressividade da doença. Segundo a *American Cancer Society*[2], as mulheres com recetores positivos hormonais tendem a ter um melhor prognóstico a curto prazo, mas esses cancros, por vezes, podem recidivar muitos anos após o tratamento. Na amostra analisada, 11,81% das pacientes tem recetores de estrogénio negativos, 20,39%

[1]Consultado a 5 de Maio de 2021, retirado de *https://www.ligacontracancro.pt/cancro-da-mama/*
[2]Consultado a 7 de Maio de 2021, retirado de *https://www.cancer.org/cancer/breast-cancer.html*

tem recetores de progesterona negativos e 44,19% das pacientes tem cancro do subtipo HER2+. As taxas de valores em falta são 13,77%, 17,89% e 25,76%, respetivamente. O cancro triplo negativo representa 3,76% dos cancros da amostra. Este é um cancro que é negativo para os recetores de estrogénio e progesterona e que não apresentam uma sobre-expressão da proteína HER2, no entanto este cresce e dissemina-se com uma maior rapidez em comparação com os outros tipos de cancro de mama. Segundo Rietjens et al.(2012), os trabalhos realizados citam os subtipos moleculares HER2 positivo e triplo negativo como tendo maiores índices de recidiva. Nesta amostra, 47,62% das pacientes com subtipo triplo negativo apresentam recidiva do cancro, no entanto, as percentagens de recidiva no subtipo HER2 são semelhantes em pacientes com resultado positivo e negativo (14,98% e 14,29%, respetivamente). Como indicado por de Oliveira e da Silva (2011), a determinação da atividade proliferativa da neoplasia, em especial o índice de imunomarcação com o anticorpo ki-67, é bastante pertinente para um correto prognóstico. Do total das pacientes em estudo, 31,84% têm ki-67 alto e 43,83% não têm informação disponível. Nesta amostra verifica-se que as percentagens de recidiva são mais elevadas em pacientes com recetores hormonais positivos, triplo negativo e com alto índice de proliferação ki-67.

Estão disponíveis duas variáveis longitudinais que serão descritas em seguida.

Segundo Guadagni, et al. (2001) *carcinoembryonic antigen* (CEA) é um marcador tumoral identificado e caracterizado no momento do diagnóstico. Este marcador é utilizado no acompanhemento pós-cirúrgico visto que a literatura sugere que este pode ajudar a identificar precocemente uma recidiva e a monitorizar a resposta ao tratamento. Contudo, alguns estudos referem que o CEA apresenta uma baixa sensibilidade em doenças precoses e avançadas comparativamente com o marcador tumoral *Carcinoma Antigen 15-3*(CA15-3). Apesar de a dispobilização do CA15-3 ter diminuido o valor do marcador CEA este continua a ser muito utilizado no acompanhemento do cancro da mama. Estes marcadores são medidos ao longo do tempo uma vez que valores acima dos valores de referência podem indicar uma possível recorrência do cancro. Considera-se o valor de referência de 5,0 ng/ml para o marcador tumoral CEA e de 37 U/ml para o marcador tumoral CA15-3.

Em relação ao marcador CEA há informação para 531 pacientes o que perfaz um total de 550 pacientes elegíveis para as análises, visto que 19 mulheres têm cancro bilateral e serão, como já mencionado, analisadas de forma independente. Há 4166 medidas do marcador tumoral CEA e o número de medições por paciente varia entre 1 e 23. O valor mediano é 7 medições por paciente. A Figura 4.10 apresenta a distribuição do número de medições por paciente no marcador tumoral CEA. Nesta amostra há 82 recidivas do cancro da mama. Relativamente às medidas do marcador após recidiva (Figura 4.11), há em média 5 medições por paciente sendo o mínimo 1 medida, o máximo 22 e a mediana 4 medidas. Se as observações após a recidiva não forem consideradas, há um total de 3853 observações e 547 pacientes das quais 79 tem recidiva.

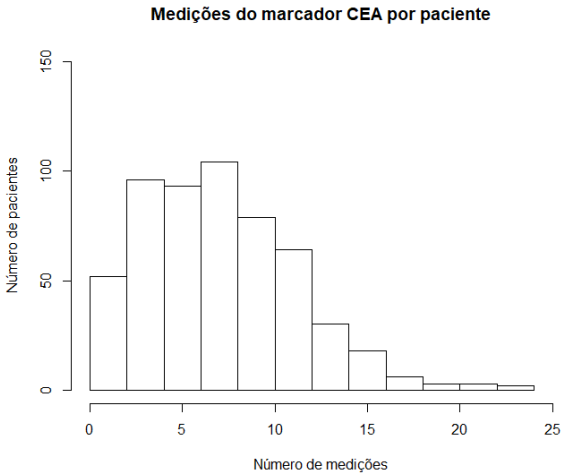

Figura 4.10: Medidas por paciente: CEA

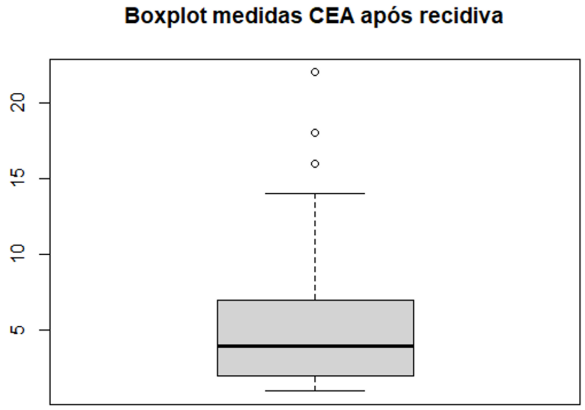

Figura 4.11: *Boxplot* medidas após recidiva: CEA

Relativamente ao marcador CA15-3 há informação para 533 pacientes o que perfaz um total de 551 pacientes selecionadas para as análises, visto que 18 mulheres têm cancro bilateral e serão, como já mencionado, analisadas de forma independente. Há 5162 medidas do marcador tumoral CA15-3, e o número de medições por paciente varia entre 1 e 48. O valor mediano é 8 medições por paciente. A Figura 4.12 apresenta a distribuição do número de medições por paciente no marcador tumoral CA15-3. Nesta amostra há 83 recidivas do cancro da mama. Relativamente às medidas do marcador após recidiva (Figura 4.13), há em média 10,25 medições por paciente sendo o mínimo 1 medida, o máximo 41 e a mediana 7 medidas. Se as observações após a recidiva não forem consideradas, há um total de 4418 observações e 550 pacientes das quais 82 tem recidiva.

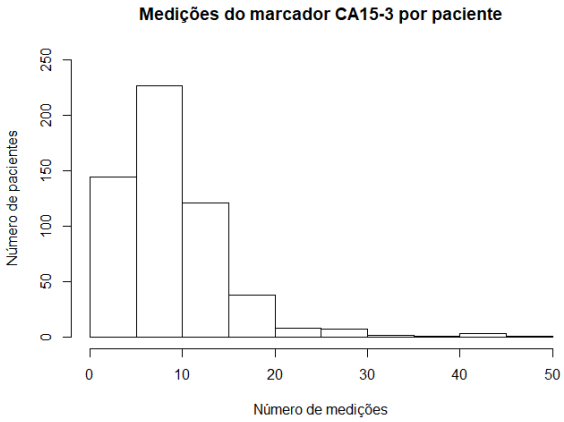

Figura 4.12: Medidas por paciente: CA15-3

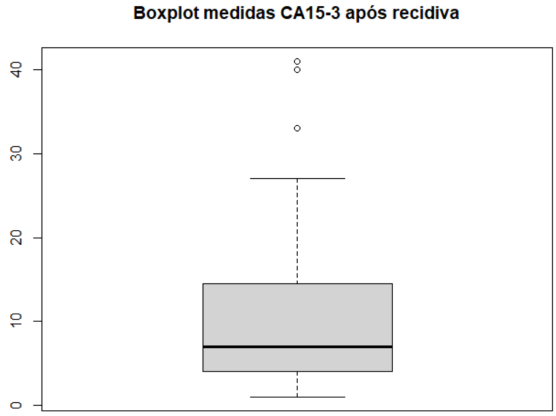

Figura 4.13: *Boxplot* medidas após recidiva: CA15-3

4.2 Análise longitudinal

4.2.1 Marcador tumoral CEA

A base de dados dispõe de 550 pacientes com diagnóstico de cancro da mama e um total de 4166 medidas do marcador tumoral CEA. Nesta base de dados há 82 recidivas do cancro da mama. Existem medições do marcador tumoral antes do diagnóstico do cancro da mama, mas estas observações não serão utilizadas nas futuras análises. Assim, há 548 pacientes em análise e 4101 medições do marcador tumoral. As duas pacientes que deixaram de estar presentes na base de dados tinham apenas uma medição do marcador antes do diagnóstico e provavelmente foram seguidas noutro hospital. Esta base de dados é a base de dados mais completa e será considerada a base de dados total (base de dados TT, em que a primeira letra se refere à população - Todas as pacientes, e a segunda letra refere-se aos dados - Todos os dados, ou seja, antes e após a recidiva). Como o evento de

interesse é a recidiva e existem pacientes que após o evento de interesse continuam a ser seguidas nas consultas, existem observações após a recidiva. Desta forma, será necessário considerar duas bases de dados distintas: a base de dados total apresentada anteriormente e uma sub-base de dados com todas as pacientes, mas que apenas contém informação das observações até recidiva (sub-base de dados TR). Esta sub-base de dados tem 545 pacientes, 3788 observações e 79 pacientes com recidivas. As pacientes que deixaram de estar presente nesta sub-base de dados tiveram recidiva antes de serem seguidas no Hospital de Braga e, desta forma, não têm observações para constar nesta sub-base de dados. Como variável tempo, considera-se o tempo desde o diagnóstico, que consiste na diferença entre a data de cada medição e a data de diagnóstico do cancro da mama.

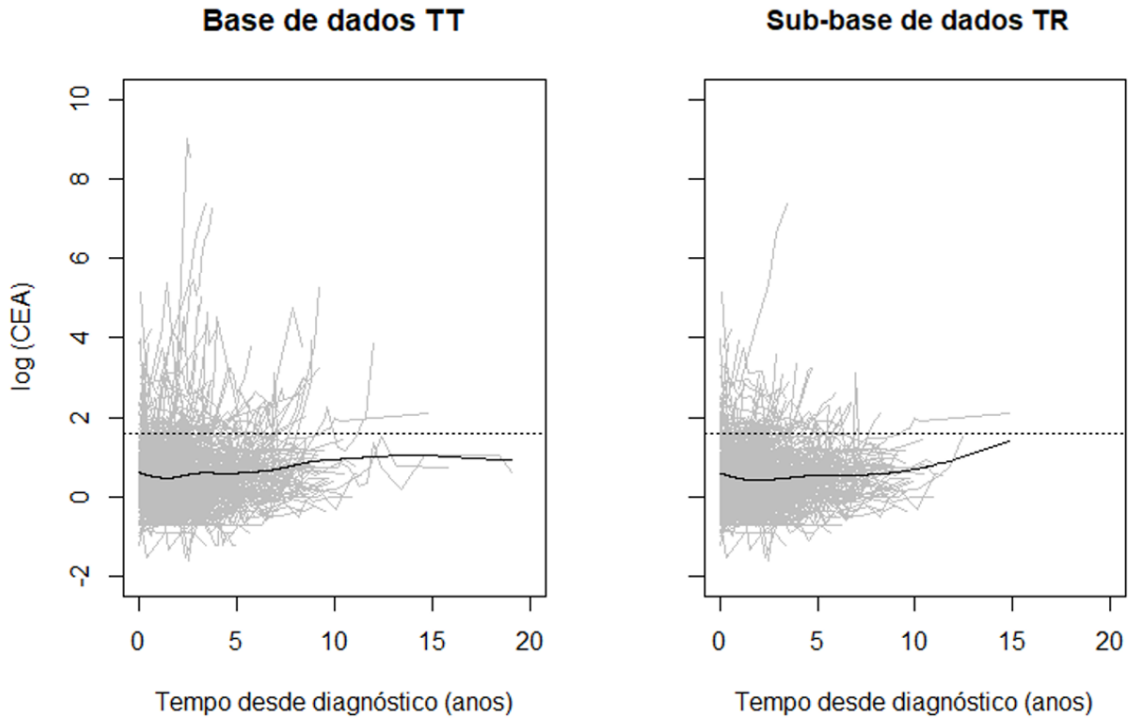

Figura 4.14: Progressões médias base de dados TT e sub-base de dados TR: CEA

A Figura 4.14 apresenta as progressões das pacientes das bases de dados referidas anteriormente sendo estudado como tempo zero o momento do diagnóstico e considerando-se os valores do marcador na escala logarítmica. A linha a tracejado é o valor de referência log (5) e a linha preta é o *smoth spline*, ou seja, o comportamento médio da progressão tendo em conta o total de pacientes em análise em cada base de dados. Através da análise das duas representações gráficas apresentadas percebe-se que ao não se considerar as observações após recidiva (sub-base de dados TR), perdem-se observações com valores do marcador mais elevados. Observa-se que na representação gráfica da base de dados TT há mais observações nos primeiros 10 anos com valores logaritmicos do marcador superiores a 3 que desaparecem na representação gráfica da sub-base de dados TR. Isto comprova que

valores mais elevados do marcador tumoral indicam um quadro clínico desfavorável e uma possível recidiva associada. Apesar do comportamento médio não ultrapassar em nenhum momento o valor de referência é bem visível que há muitos pacientes com valores bastante mais elevados que o valor recomendado. Para melhor entender a progressão das pacientes com recidivas, analisam-se em seguida duas sub-bases de dados que consideram apenas pacientes que tiveram recidiva do cancro da mama. Uma sub-base de dados contém todas as observações das pacientes com recidiva (sub-base de dados RT) e a outra sub-base de dados apenas contém informação das observações até à recidiva (sub-base de dados RR). A sub-base de dados RT tem 82 pacientes e 717 observações enquanto a sub-base de dados RR tem 79 pacientes e 406 observações. A Tabela 4.4, resume as bases de dados apresentadas e as respetivas descrições.

Tabela 4.4: Descrição das bases de dados

Base de Dados	**Descrição**
Base de dados TT	Base de dados com todos os pacientes e todas as observações
Sub-base de dados TR	Base de dados com todos os pacientes e observações até recidiva
Sub-base de dados RT	Base de dados com pacientes com recidiva e todas as observações
Sub-base de dados RR	Base de dados com pacientes com recidiva e observações até recidiva

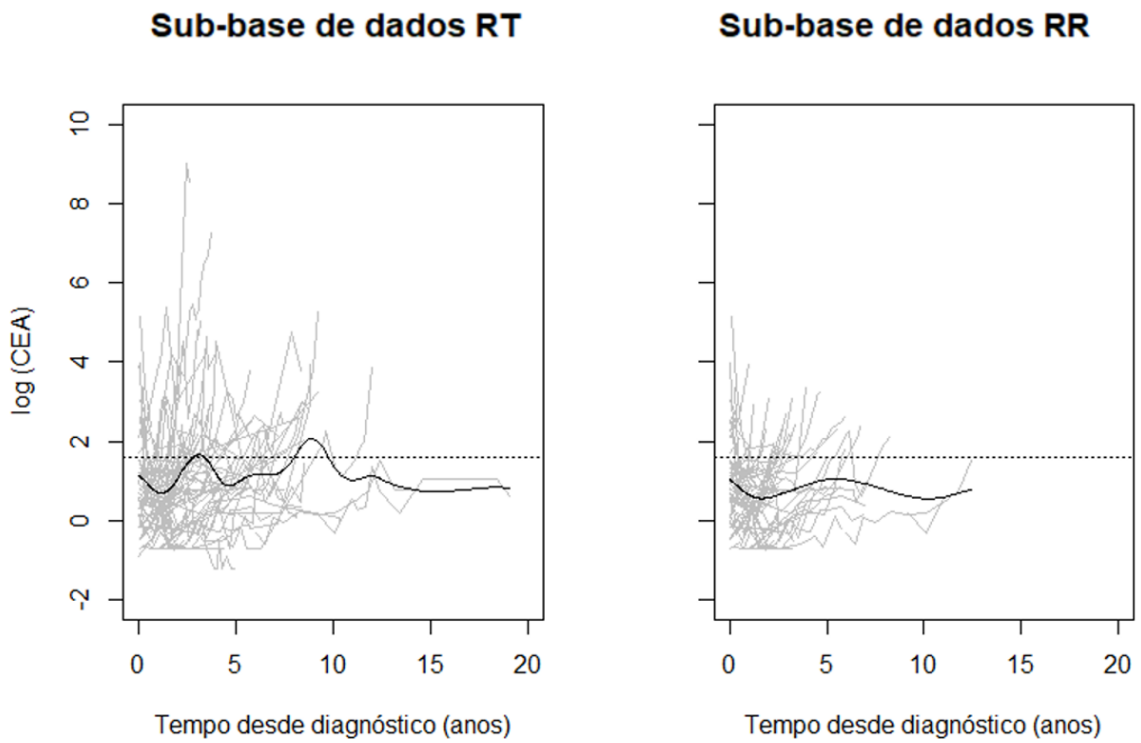

Figura 4.15: Progressões médias sub-base de dados RT e sub-base de dados RR: CEA

Analisando-se a Figura 4.15 observa-se que a perda de observações com valores mais elevados do marcador é ainda mais evidente nestes dois gráficos dada a maior facilidade de interpretação devido à diminuição de pacientes em análise. Salienta-se ainda que o comportamento médio da sub-base de dados com todas as observações ultrapassa o valor de referência em dois momentos distintos. A Figura 4.16 apresenta apenas os comportamentos médios considerando as quatro bases de dados apresentadas. Nas sub-bases de dados com observações até recidiva nota-se uma subida abrupta após os 10 anos, as subidas dos valores do marcador indicam uma piora do estado clínico e está associada à recidiva da doença. As sub-base de dados com todas as pacientes tem um maior comprimento pois existirão pacientes sem recidiva que foram acompanhadas por mais anos.

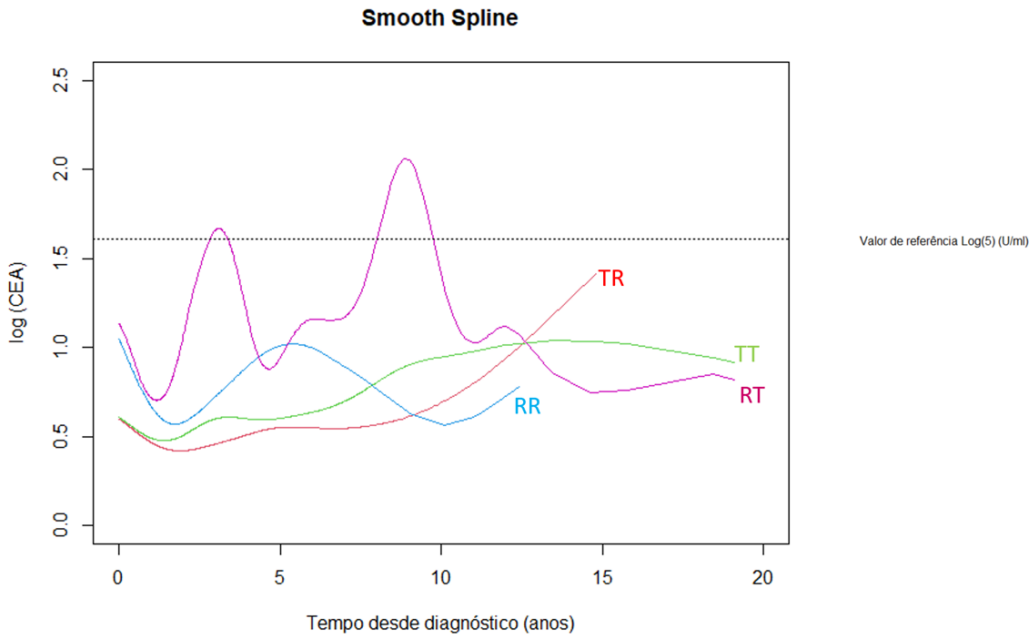

Figura 4.16: Comparação dos Smooth Splines: CEA

Considerando-se como momento zero a recidiva (Figura 4.17) e analisando-se apenas as pacientes com recidiva do cancro, verifica-se que logo após a recidiva os valores do marcador parecem estabilizar e depois há um aumento acentuado. A estabilização dos valores pode ser resultante da aplicação do tratamento logo após a deteção da recidiva. Visto que há uma grande percentagem de pacientes que morrem após a recidiva (cerca de 64%), os valores após os 3 anos podem induzir em erro uma vez que apenas se estão a observar as pacientes com melhores resultados.

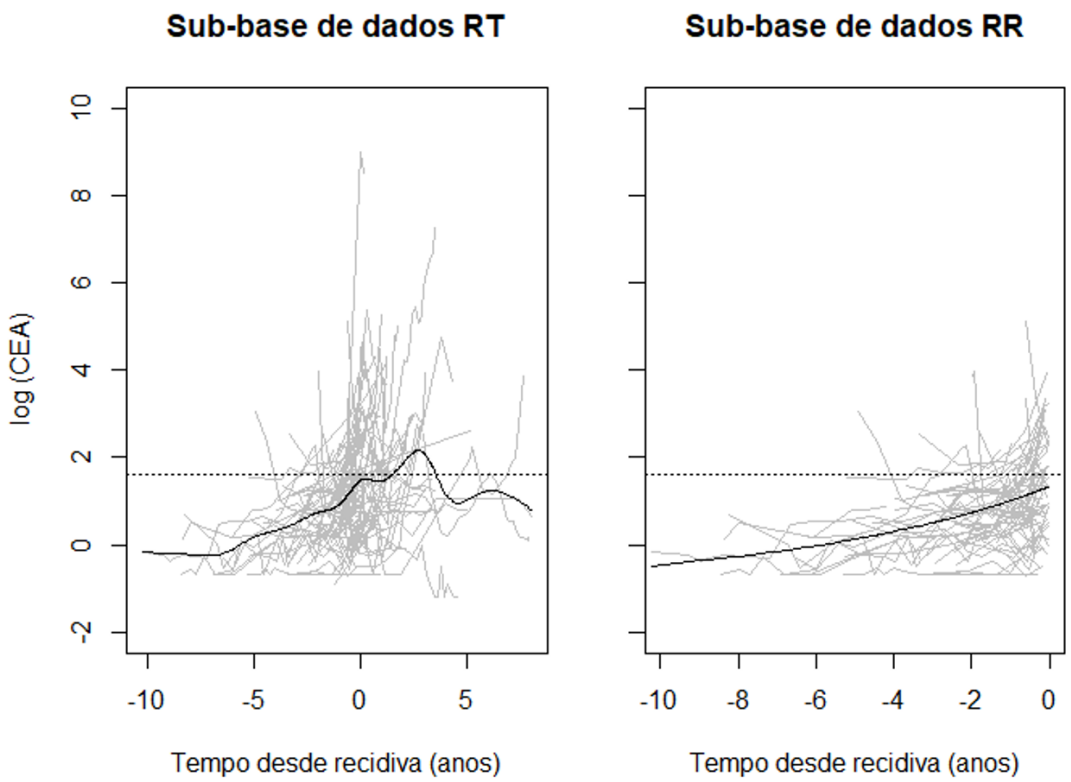

Figura 4.17: Progressões médias sub-base de dados RT e sub-base de dados RR, tempo desde recidiva: CEA

Começou-se por analisar a estrutura de correlação, assim construíram-se o variograma empírico e os variogramas teóricos tendo em consideração três diferentes estruturas de correlação: estrutura só com efeitos aleatórios, estrutura de correlação exponencial e estrutura de correlação gaussiana. Para a estimação destes variogramas utilizaram-se os modelos saturados com um total de 16 variáveis explicativas.

A Figura 4.18 apresenta os variogramas tendo em conta a base de dados total e a sub-base de dados com todas as pacientes e observações até recidiva. A linha a preto representa o variograma empírico, a linha a azul o variograma considerando apenas efeitos aleatórios, a linha a verde o variograma considerando a estrutura de correlação exponencial e a linha a vermelho o variograma considerando a estrutura de correlação gaussiana.

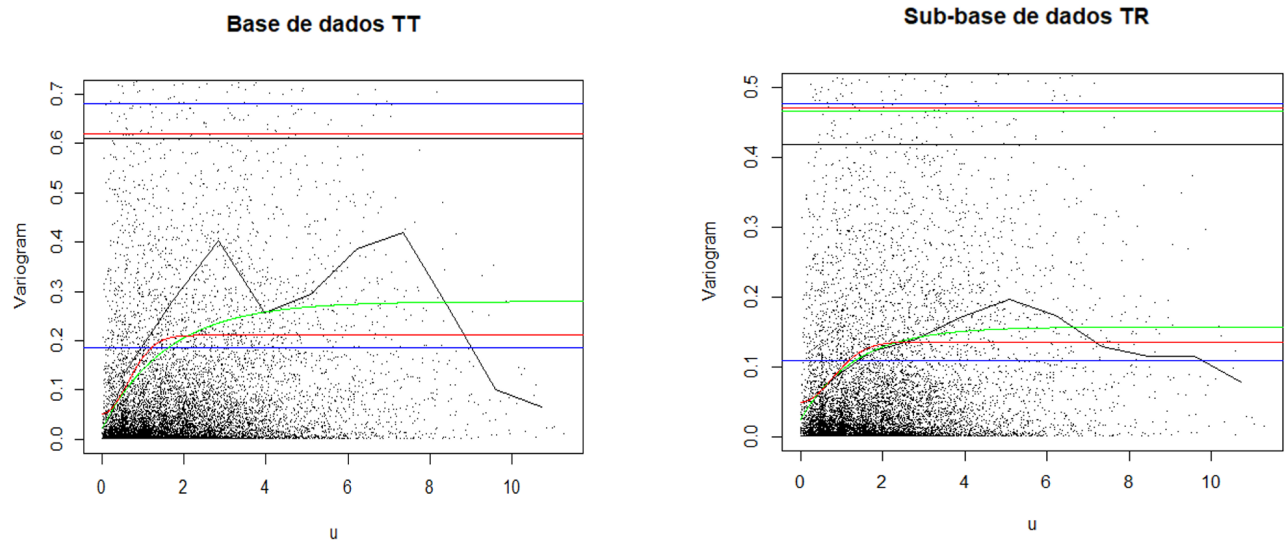

Figura 4.18: Variogramas teóricos e empírico: CEA, base de dados TT e sub-base de dados TR

Tabela 4.5: Estrururas de correlação CEA: base de dados TT e sub-base de dados TR

Estrutura de correlação	Base de dados TT				Sub-base de dados TR			
	AIC	logLik	número de observações	número parâmetros	AIC	logLik	número de observações	número parâmetros
Só com efeito aleatório	4626,64	-2293,32	2937	20	3000,54	-1480,27	2724	20
Exponencial	**3511,22**	**-1733,61**	2937	22	**2451,92**	**-1203,96**	2724	22
Gaussiana	3653,14	-1804,57	2937	22	2523,26	-1239,63	2724	22

Analisando-se a estrutura de correlação com base na análise gráfica e nos resultados de AIC e Log Likelihood (Tabela 4.5), concluí-se que em ambas as bases de dados a estrutura de correlação exponencial é a que melhor de adequa.

Assim, a Tabela 4.6 apresenta os resultados obtidos para a estimação dos coeficientes dos modelos com estrutura de correlação exponencial. Em ambos os modelos a progressão varia ao longo do tempo e, dada a significância da interação do tempo com a recidiva, a progressão de pacientes com e sem recidiva são diferentes. Pacientes com recidiva têm uma progressão mais acelerada e valores iniciais na escala logarítmica do marcador tumoral mais elevados. Pode-se inferir que a idade ao diagnóstico afeta o valor do logaritmo do marcador aumentando 0,017 por ano de idade no diagnóstico. Considerando-se que uma variável é significativa quando apresenta um valor de prova (p-valor) inferior a 0,05, a variável Tumor primário poderia ser considerada marginalmente significativa quando se utiliza para a estimação dos coeficientes a base de dados TT.

Tabela 4.6: Parâmetros estimados do modelo saturado com estrutura de correlação exponencial CEA: TT e TR

Estrutura de correlação exponencial	**Base de dados TT**		**Sub-base de dados TR**	
	β	**p-valor**	β	**p-valor**
Intercept	**-0,835**	**0,026**	-0,625	0,063
Tempo desde diagnóstico (anos)	**0,026**	**0,002**	**0,029**	**<0,0001**
Tempo desde diagnóstico:Recidiva	**0,181**	**<0,0001**	**0,149**	**<0,0001**
Idade ao diagnóstico (continua)	**0,017**	**<0,0001**	**0,017**	**<0,0001**
Recidiva (sim)	**0,323**	**0,008**	**0,220**	**0,043**
Bilateral (sim)	-0,048	0,756	-0,053	0,702
Estadio (III ou IV)	0,129	0,631	0,026	0,914
Grau (G3)	0,086	0,390	0,011	0,901
Menopausa (sim)	-0,204	0,082	-0,152	0,148
Mama (esquerda)	0,083	0,236	0,043	0,497
Presença de carcinoma associado (sim)	0,111	0,121	0,075	0,240
Invasão vascular linfática (sim)	0,007	0,937	0,043	0,584
Invasão vascular venosa (sim)	0,194	0,265	0,083	0,595
Recetor de estrogénio (positivo)	0,162	0,279	0,010	0,943
Recetor de progesterone (positivo)	-0,107	0,312	0,020	0,831
Triplo negativo (sim)	-0,379	0,086	-0,188	0,351
Tumor primário (T3 ou T4 ou Tx)	**0,344**	**0,054**	0,156	0,325
Nódulos linfáticos (Nx ou N0 ou N1)	0,167	0,531	0,000	0,999

Fazendo a mesma análise, mas agora considerando apenas as pacientes com recidiva da doença, consegue-se perceber que as estruturas de correlação temporal se mantêm. A Figura 4.19 e a Tabela 4.7 comprovam esta afirmação. Os modelos que estão na base da estimação destes variogramas tem em consideração 16 variáveis explicativas, no entanto a variável recidiva foi substituída pela variável tipo de recidiva para se conseguir concluir sobre as diferenças existentes entre os três tipos de recidivas.

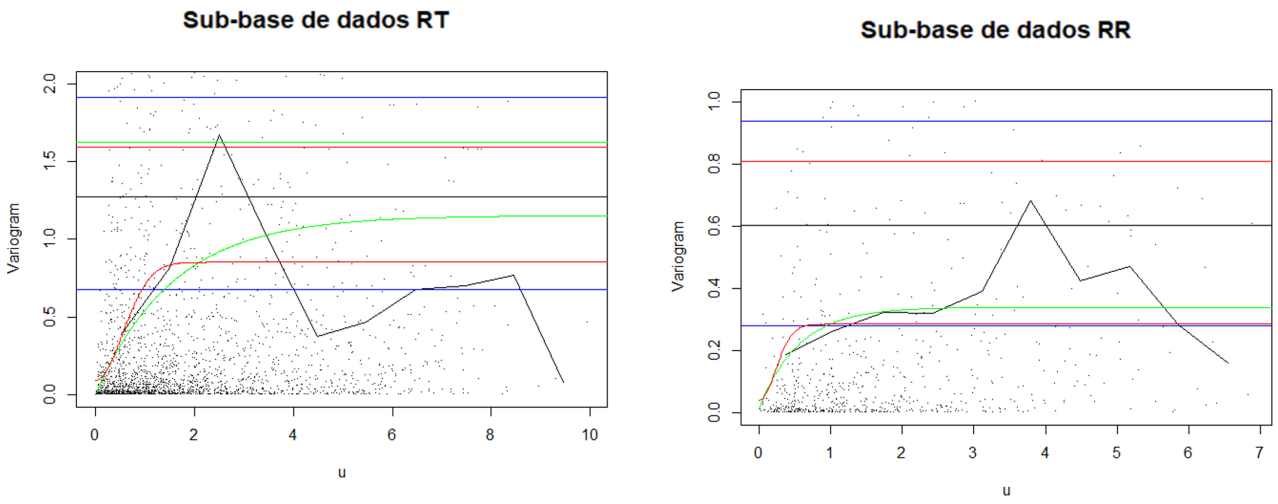

Figura 4.19: Variogramas teóricos e empírico: CEA, sub-base de dados RT e sub-base de dados RR

Tabela 4.7: Estrururas de correlação CEA: base de dados RT e sub-base de dados RR

Estrutura de correlação	Base de dados TT				Sub-base de dados TR			
	AIC	logLik	número de observações	número parâmetros	AIC	logLik	número de observações	número parâmetros
Só com efeito aleatório	1261,19	-610,60	447	20	546,36	-253,18	248	20
Exponencial	**964,47**	**-460,23**	447	22	**505,10**	**-230,55**	248	22
Gaussiana	1001,22	-478,61	447	22	511,99	-234,00	248	22

Analisando os resultados apresentados na Tabela 4.8 concluí-se que a progressão varia ao longo do tempo nos dois modelos. Pacientes com subtipo triplo negativo apresentam valores inicias mais baixos em ambos os modelos e na sub-base de dados com todas as observações os valores iniciais do marcador para pacientes com carcinoma associado são mais elevados comparativamente com os valores iniciais de pacientes sem carcinoma associado. No modelo utilizando a sub-base de dados RT a variável estadio é marginalmente significativa.

Tabela 4.8: Parâmetros estimados do modelo saturado com estrutura de correlação exponencial CEA: RT e RR

Estrutura de correlação exponencial	**Base de dados RT**		**Sub-base de dados RR**	
	β	**p-valor**	β	**p-valor**
Intercept	-0,098	0,946	0,631	0,577
Tempo desde diagnóstico (anos)	**0,211**	**<0,0001**	**0,187**	**<0,0001**
Presença de carcinoma associado (sim)	**0,826**	**0,043**	0,298	0,339
Triplo negativo (sim)	**-2,449**	**0,011**	**-1,511**	**0,046**
Idade ao diagnóstico (continua)	0,002	0,921	0,009	0,557
Bilateral (sim)	-1,159	0,066	-0,714	0,161
Estadio (III ou IV)	**1,751**	**0,054**	0,827	0,261
Grau (G3)	0,411	0,325	-0,175	0,591
Menopausa (sim)	0,040	0,948	0,152	0,752
Tipo de recidiva (metastizada)	-0,065	0,899	0,128	0,755
Tipo de recidiva (local e metastizada)	0,661	0,429	0,203	0,752
Mama (esquerda)	0,520	0,115	0,265	0,296
Invasão vascular linfática (sim)	-0,237	0,559	0,076	0,811
Invasão vascular venosa (sim)	0,087	0,881	-0,146	0,741
Recetor de estrogénio (positivo)	-0,537	0,507	-0,774	0,208
Recetor de progesterone (positivo)	-0,916	0,066	-0,626	0,104
Tumor primário (T3 ou T4 ou Tx)	0,264	0,591	0,034	0,926
Nódulos linfáticos (Nx ou N0 ou N1)	1,046	0,222	0,032	0,964

Em conslusão, o subtipo triplo negativo e a presença de carcinoma associado na fase de diagnóstico parecem influenciar os valores do marcador quando as pacientes têm recidiva não sendo possível detetar diferenças significativas nos valores dos marcadores quando se consideram todas as pacientes em estudo.

4.2.1.1 Modelo final com estrutura de correlação exponencial para o marcador tumoral CEA

Considerando a base de dados total e a estrutura de correlação exponencial obteve-se o modelo apresentado na Tabela 4.9.

Tabela 4.9: Modelo final CEA com estrutura de correlação exponencial

Estrutura de correlação exponencial	**Base de dados TT**	
	$\boldsymbol{\beta}$	**p-valor**
Intercept	-0,412	0,006
Tempo desde diagnóstico (anos)	0,023	0,009
Tempo desde diagnóstico:Recidiva	0,145	<0,0001
Idade ao diagnóstico (continua)	0,014	<0,0001
Recidiva (sim)	0,392	0,001
Triplo negativo (sim)	-0,383	0,025
Tumor primário (T3 ou T4 ou Tx)	0,364	0,003
$\hat{\nu}^2$	0,329	
$\hat{\sigma}^2$	0,330	
$\hat{\phi}$	2,185	
$\hat{\tau}^2$	0,024	
AIC	4223,923	
Log Likelohood	-2100,962	

Ao analisar os resultados concluí-se que a progressão varia ao longo do tempo aumentando 0,023 do valor do logaritmo do marcador tumoral CEA por ano. Pacientes com recidiva apresentam um incremento de 0,392 no ponto inicial da progressão média dos valores do logaritmo do marcador tumoral e uma progressão mais acelerada, cerca de mais 0,145 por ano em comparação a mulheres sem recidiva. O valor 0,014, para idade ao diagnóstico contínuo, significa que há um aumento desse valor por ano de idade no diagnóstico. As variáveis triplo negativo e tumor primário têm um efeito estatisticamente significativo na progressão média deste marcador tumoral, havendo um incremento de 0,364 no ponto inicial da progressão média dos valores do logaritmo do marcador tumoral em pacientes com tumor primário T3 ou T4 ou Tx em comparação com pacientes com tumor primário T1 ou T2 e uma diminuição de 0,383 em pacientes com subtipo triplo negativo. A estrutura de correlação que melhor representa a variabilidade dos dados é aquela que incorpora efeitos aleatórios em nível individual com $\hat{\nu}^2 = 0,329$, uma estrutura de correlação exponencial para descrever a variabilidade dentro dos pacientes com $\hat{\rho}^2 = \exp(-\frac{1}{2,185}|u|)$ e $\hat{\sigma}^2 = 0,330$ e um erro de medição com variância $\hat{\tau}^2 = 0,024$.

4.2.1.2 Análise de diagnóstico

Na análise de diagnóstico validam-se os pressupostos do modelo. A adequabilidade da estrutura de correlação escolhida já foi analisada anteriormente. As Figuras 4.20 e 4.21 apresentam o gráfico dos resíduos específicos do sujeito *versus* os valores ajustados e um gráfico Q-Q normal dos resíduos específicos do sujeito, respetivamente. Observando estes gráficos, verifica-se que na Figura 4.20 para valores ajustados mais elevados os resíduos padronizados aumentam, não estando em torno de zero. Na Figura 4.21 existe algum desvio em relação à reta. No entanto, os pressupostos do modelo longitudinal, relativos à homogeneidade das variâncias dos erros de medição e à sua normalidade, não parecem passíveis de rejeição.

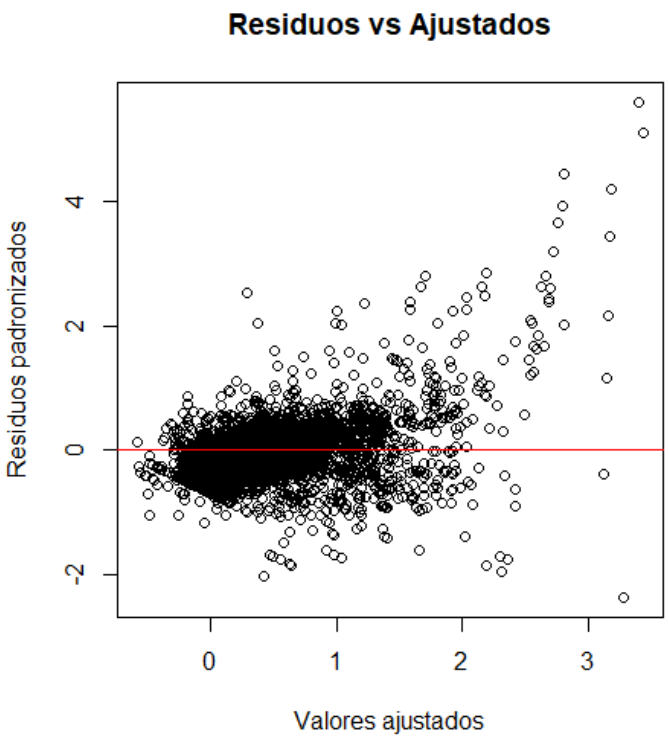

Figura 4.20: Resíduos específicos do sujeito *versus* os valores ajustados: CEA

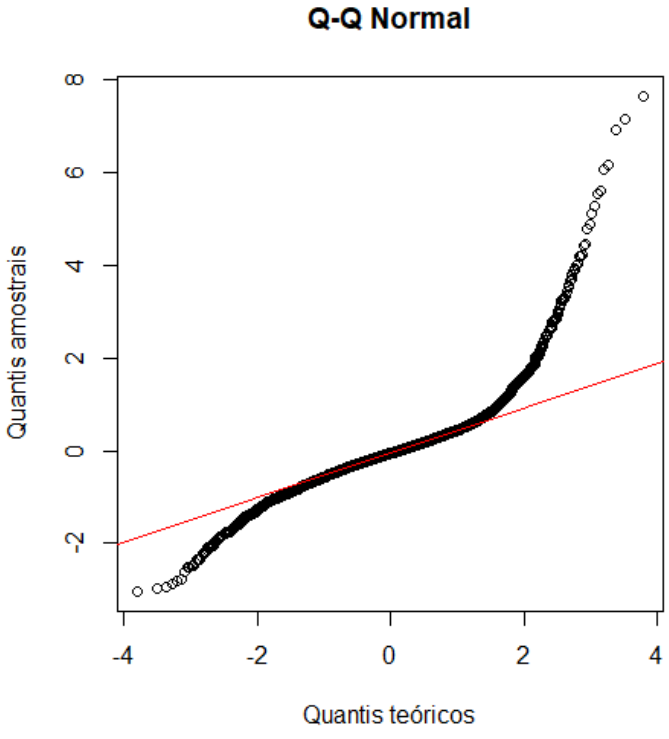

Figura 4.21: Q-Q dos resíduos específicos do sujeito: CEA

4.2.2 Marcador tumoral CA15-3

A base de dados dispõe de 551 pacientes com diagnóstico de cancro da mama e um total de 5162 medidas do marcador tumoral CA15-3. Nesta base de dados há 83 recidivas do cancro da mama. Existem medições do marcador tumoral antes do diagnóstico do cancro da mama, mas estas observações não serão utilizadas nas futuras análises. Assim, há 550 pacientes em análise e 5111 medições do marcador tumoral. A paciente que deixou de estar na base de dados tinha apenas uma medição do marcador antes do diagnóstico e provavelmente foi seguida noutro hospital. Esta base de dados é a base de dados mais completa e será considerada a base de dados total (base de dados TT). Como o evento de interesse é a recidiva e existem pacientes que após o evento de interesse continuam a ser seguidas nas consultas, existem observações após a recidiva. Deste forma, será necessário considerar duas bases de dados distintas, a base de dados total apresentada anteriormente e uma sub-base de dados com todas as pacientes, mas que apenas contém informação das observações até recidiva (sub-base de dados TR). Esta sub-base de dados tem 549 pacientes, 4367 observações e 82 pacientes com recidivas. A paciente que deixou de estar presente nesta sub-base de dados teve recidiva antes de ser seguida no Hospital de Braga e, desta forma, não tem observações para constar nesta sub-base de dados. Como variável tempo, considera-se o tempo desde o diagnóstico que consiste na diferença entre a data de cada medição do marcador tumoral e a data de diagnóstico do cancro da mama.

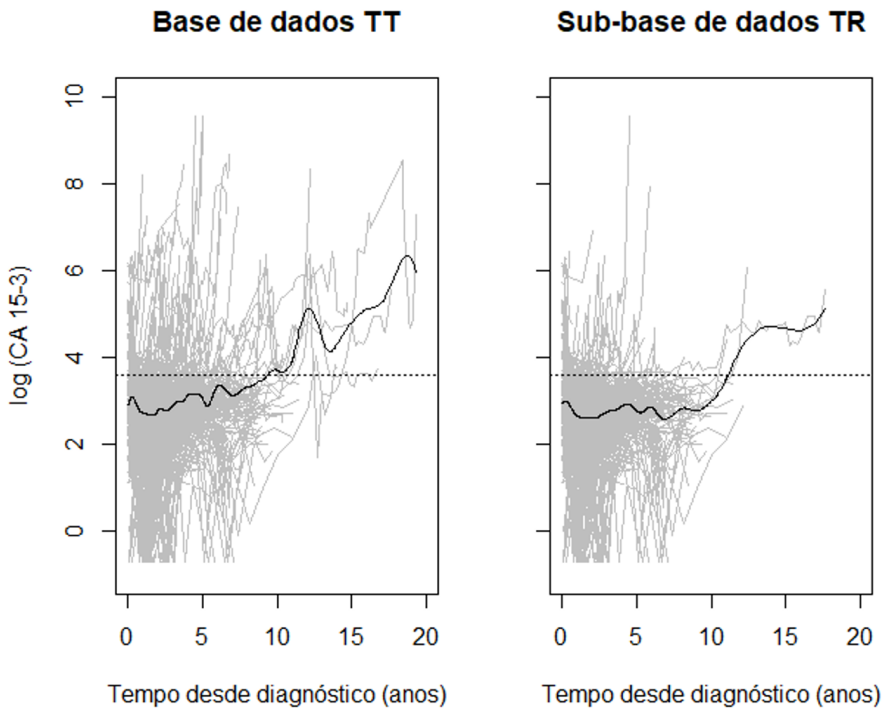

Figura 4.22: Progressões médias base de dados TT e sub-base de dados TR: CA15-3

A Figura 4.22 apresenta as progressões das pacientes das bases de dados referidas anteriormente estudando-se como tempo zero o momento do diagnóstico e considerando-se os valores do marcador na escala logarítmica. A linha a tracejado é o valor de referência log (37) e a linha preta é o *smooth spline*, ou seja, o comportamento médio da progressão tendo em conta o total de pacientes em análise em cada base de dados. Através da análise das duas representações gráficas apresentadas percebe-se que ao não se considerar as observações após recidiva, perdem-se observações com valores do marcador mais elevados. Na representação gráfica da base de dados TT existem muitos valores do marcador acima do valor de referência que desaparecem quando não se consideram as observações após recidiva. Isto comprova que valores mais elevados do marcador tumoral indicam um quadro clínico desfavorável e uma possível recidiva associada. O comportamento médio ultrapassa o valor de referência após os 10 anos, sensivelmente, e nota-se uma perda significativa de pacientes a partir desse ponto, assim os resultados após os 10 anos devem ser cautelosamente analisados.

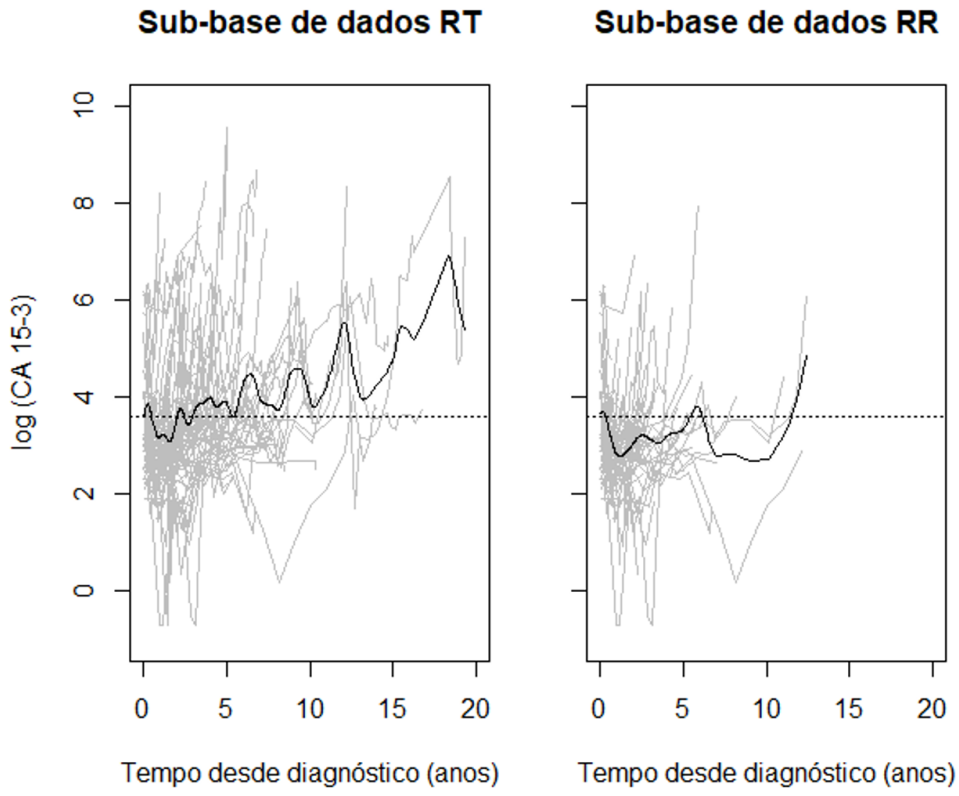

Figura 4.23: Progressões médias sub-base de dados RT e sub-base de dados RR: CA15-3

Para se entender melhor a progressão das pacientes com recidivas, analisaram-se as duas sub-bases de dados que consideram apenas pacientes que tiveram recidiva do cancro da mama. Uma sub-base de dados contém todas as observações das pacientes com recidiva (sub-base de dados RT) e a outra sub-base de dados apenas contém informação das

observações até à recidiva (sub-base de dados RR). A sub-base de dados RT tem 83 pacientes e 1244 observações enquanto a sub-base de dados RR tem 82 pacientes e 502 observações. Analisando-se a Figura 4.23 observa-se que a perda de observações com valores mais elevados do marcador é ainda mais evidente nestes dois gráficos dada a maior facilidade de interpretação devido à diminuição de pacientes em análise. Quando se analisam todas as observações o valor médio da progressão ultrapassa o valor de referência e mantém-se acima deste até aos 20 anos após diagnóstico. Quando apenas se analisam as observações até à recidiva, o valor de referência é ultrapassado em dois momentos distinto e deve-se ter especial atenção na interpretação dos valores após os 5 anos porque há poucas pacientes a serem analisadas.

A Figura 4.24 apresenta apenas os comportamentos médios considerando as quatro bases de dados apresentadas.

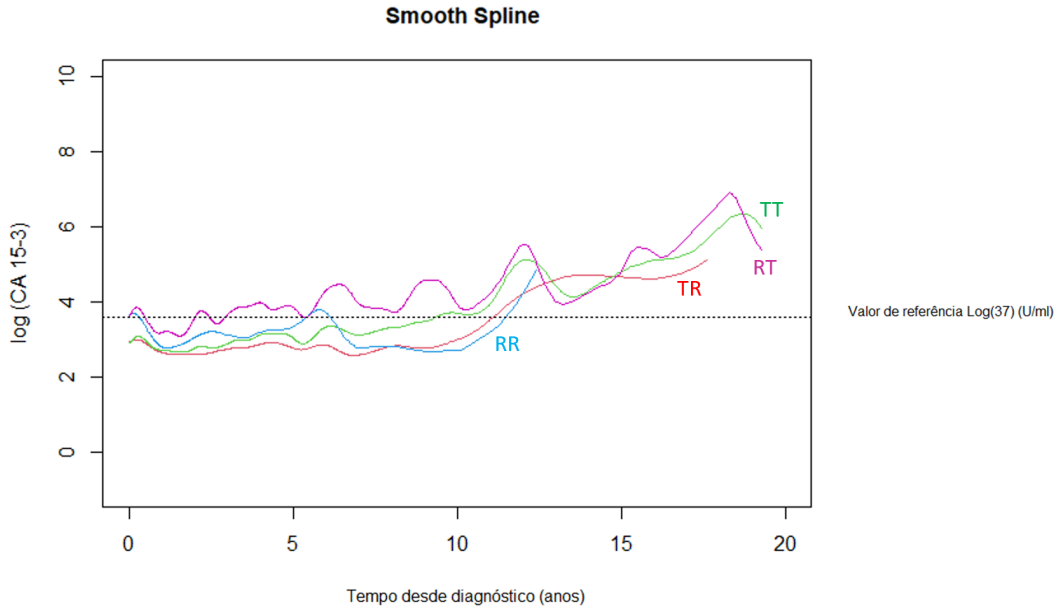

Figura 4.24: Comparação dos Smooth Splines: CA15-3

Todos os comportamentos médios ultrapassam o valor de referência e há uma subida acentuada em todos por volta dos 10 anos. A progressão considerando-se todas as observações das pacientes com recidiva diferencia-se das demais uma vez que ultrapassa o valor de referência muito mais cedo que as restantes e mantém-se sempre acima deste.

Considerando-se como momento zero a recidiva (Figura 4.25) e analisando-se apenas as pacientes com recidiva do cancro, verifica-se que logo após a recidiva os valores do marcador parecem estabilizar e depois há um aumento acentuado mantendo-se sempre acima do valor de referência. A estabilização dos valores pode ser resultante da aplicação do tratamento logo após a deteção da recidiva. Visto que há uma grande percentagem

de pacientes que morrem após a recidiva (cerca de 64%) os valores após os 3 anos podem induzir em erro uma vez que se está apenas a observar as pacientes com melhores resultados.

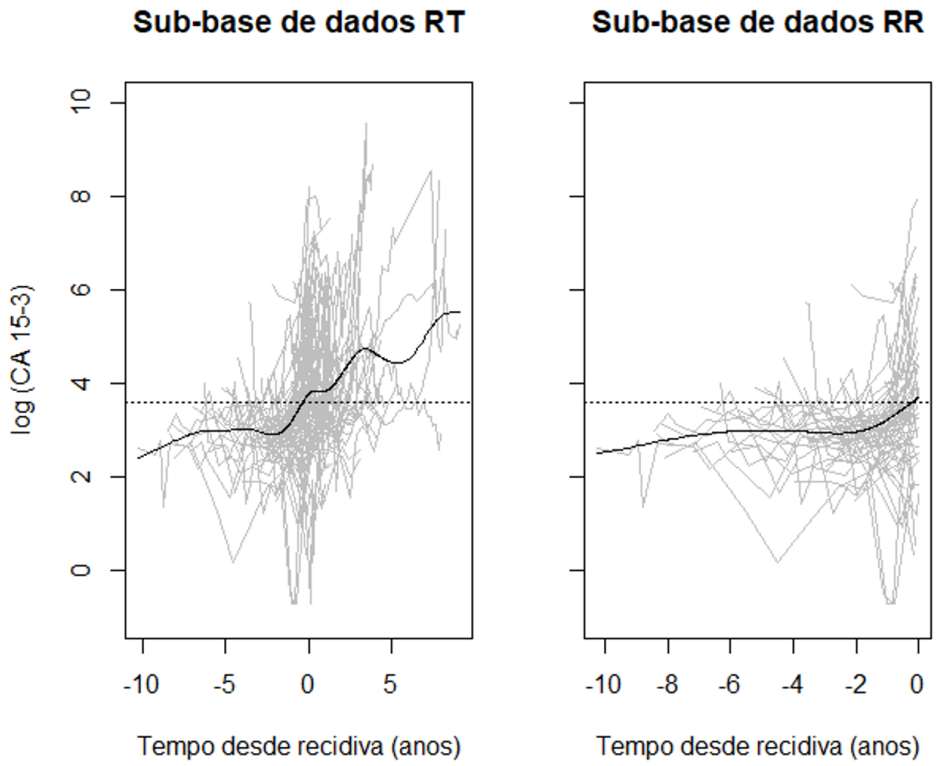

Figura 4.25: Progressões médias sub-base de dados RT e sub-base de dados RR, tempo desde recidiva: CA15-3

Começou-se por analisar a estrutura de correlação, assim construiram-se o variograma empírico e os variogramas teóricos tendo em consideração três diferentes estruturas de correlação: estrutura só com efeitos aleatórios, estrutura de correlação exponencial e estrutura de correlação gaussiana. Para a estimação destes variogramas utilizaram-se os modelos saturados com um total de 16 variáveis explicativas. A Figura 4.26 apresenta os variogramas tendo em conta a base de dados total e a sub-base de dados com todas as pacientes e observações até recidiva. A linha a preto representa o variograma empírico, a linha a azul o variograma considerando apenas efeitos aleatórios, a linha a verde o variograma considerando a estrutura de correlação exponencial e a linha a vermelho o variograma considerando a estrutura de correlação gaussiana.

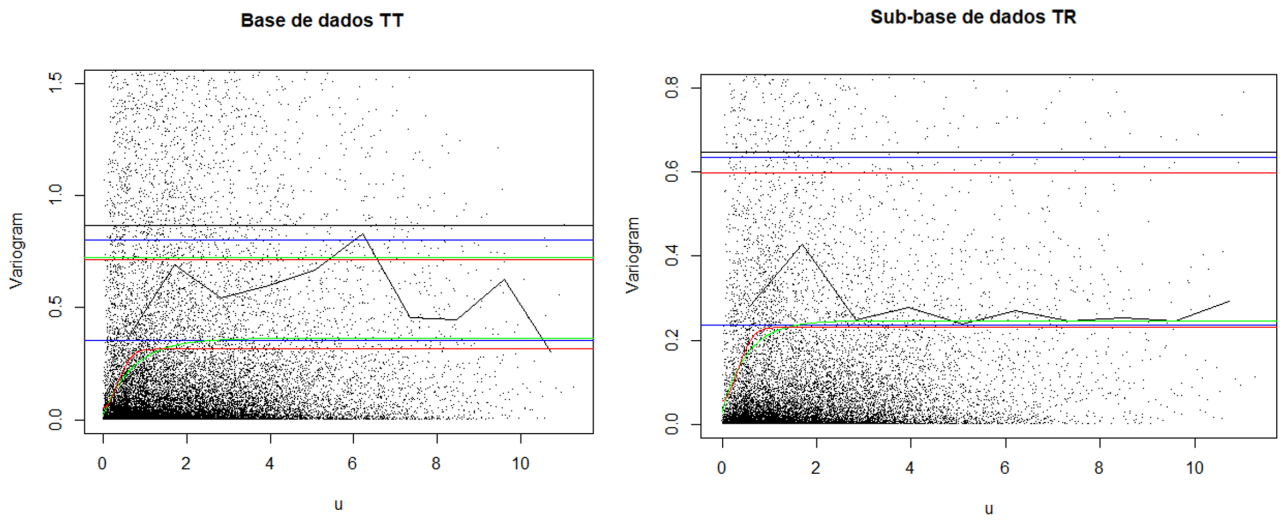

Figura 4.26: Variogramas teóricos e empírico: CA15-3, base de dados TT e sub-base de dados TR

Tabela 4.10: Estrururas de correlação CA15-3: base de dados TT e sub-base de dados TR

Estrutura de correlação	Base de dados TT				Sub-base de dados TR			
	AIC	logLik	número de observações	número parâmetros	AIC	logLik	número de observações	número parâmetros
Só com efeito aleatório	7637,18	-3798,59	3670	20	5460,98	-2710,49	3139	20
Exponencial	**5862,07**	**-2909,04**	3670	22	**4727,34**	**-2341,67**	3139	22
Gaussiana	5964,78	-2960,39	3670	22	4728,46	-2342,23	3139	22

Analisando-se a estrutura de correlação com base na análise gráfica e nos resultados de AIC e Log Likelihood (Tabela 4.10), concluí-se que em ambas as bases de dados a estrutura de correlação exponencial é a que melhor de adequa.

A Tabela 4.11 apresenta os resultados obtidos para a estimação dos coeficientes dos modelos com estrutura de correlação exponencial. Quando se tem em conta todas as observações, concluí-se que a progressão de pacientes com recidiva é mais acelerada e que pacientes sem recidiva não parecem apresentar uma variação ao longo do tempo estatisticamente significativa. Quando se consideram apenas as observações até à recidiva, a progressão não parece variar ao longo do tempo para pacientes com ou sem recidiva. Em ambos os modelos parece existir um aumento do valor de 0,011 no valor do logaritmo do marcador tumoral por cada ano de idade no diagnóstico. Pacientes com recidiva, presença de carcinoma associado e invasão vascular venosa apresentam valores iniciais da progressão média dos valores do logaritmo do marcador tumoral mais elevados.

Tabela 4.11: Parâmetros estimados do modelo saturado com estrutura de correlação exponencial CA15-3: TT e TR

Estrutura de correlação exponencial	**Base de dados TT**		**Sub-base de dados TR**	
	β	**p-valor**	β	**p-valor**
Intercept	**1,579**	**<0,0001**	**1,710**	**<0,0001**
Tempo desde diagnóstico (anos)	0,008	0,389	0,006	0,394
Tempo desde diagnóstico:Recidiva	**0,190**	**<0,0001**	0,044	0,100
Idade ao diagnóstico (continua)	**0,011**	**0,006**	**0,011**	**0,004**
Recidiva (sim)	**0,413**	**0,001**	**0,524**	**<0,0001**
Presença de carcinoma associado (sim)	**0,149**	**0,037**	**0,141**	**0,037**
Invasão vascular venosa (sim)	**0,560**	**0,001**	**0,494**	**0,003**
Bilateral (sim)	0,164	0,298	0,193	0,198
Estadio (III ou IV)	0,403	0,063	0,262	0,201
Grau (G3)	0,033	0,744	-0,020	0,828
Menopausa (sim)	-0,155	0,186	-0,122	0,268
Mama (esquerda)	-0,093	0,184	-0,107	0,105
Invasão vascular linfática (sim)	0,044	0,624	0,020	0,815
Recetor de estrogénio (positivo)	0,101	0,497	0,046	0,745
Recetor de progesterone (positivo)	-0,065	0,537	-0,035	0,727
Triplo negativo (sim)	-0,365	0,097	-0,398	0,057
Tumor primário (T3 ou T4 ou Tx)	-0,047	0,783	0,011	0,948
Nódulos linfáticos (Nx ou N0 ou N1)	0,392	0,078	0,315	0,135

A Figura 4.27 apresenta os variogramas tendo em conta a sub-base de dados com todas as observações e a sub-base de dados com observações até recidiva para as pacientes com recidiva do cancro da mama. Os modelos que estão na base da estimação destes variogramas tem em consideração 16 variáveis explicativas, no entanto a variável recidiva foi substituída pelo variável tipo de recidiva para se conseguir concluir sobre as diferenças existentes entre os três tipos de recidiva. A Tabela 4.12 apresenta os AIC e a Log Likelihood para se concluir sobre a adequabilidade de cada estrutura de correlação consideradas.

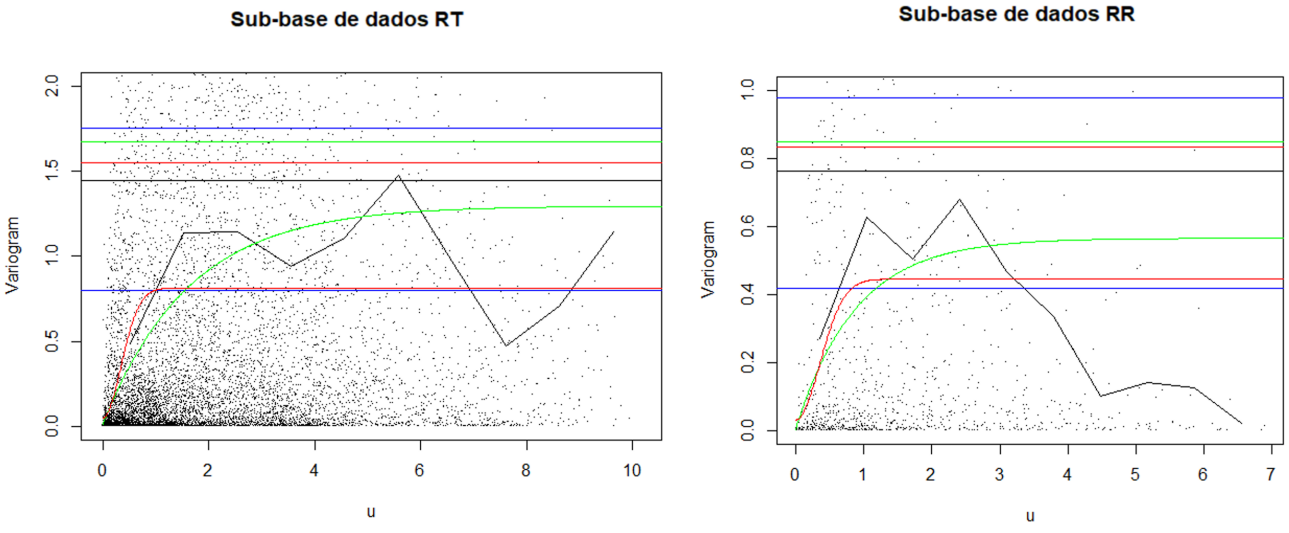

Figura 4.27: Variogramas teóricos e empírico: CA15-3, base de dados RT e sub-base de dados RR

Como se pode observar no variograma da sub-base de dados RR após o u=3 existe um número reduzido de pontos a contribuir para o variograma, dado que só 309 observações estão a ser tidas em conta o que se traduz num baixo número de consultas espaçadas mais de 3 anos, explicando a quebra abrupta existente no final do mesmo. Assim, a correlação gaussiana é sugerida pela análise do AIC (Tabela 4.12) porque o pressuposto da estabilização da correlação a partir de um determinado ponto fica condicionado e o variograma teórico da estrutura de correlação gaussiana parece acompanhar melhor esta queda final. No entanto, optou-se pela correlação exponencial, visto que pela interpretação gráfica esta parece ser adequada, ajustando bem a parte inicial onde há uma maior quantidade de pontos e ignorando a queda abrupta justificada já anteriormente.

Tabela 4.12: Estrururas de correlação CA15-3: base de dados RT e sub-base de dados RR

Estrutura de correlação	Base de dados TT				Sub-base de dados TR			
	AIC	logLik	número de observações	número parâmetros	AIC	logLik	número de observações	número parâmetros
Só com efeito aleatório	2341,49	-1150,74	826	20	756,00	-358,00	309	20
Exponencial	**1428,27**	**-692,13**	826	22	624,74	-290,37	309	22
Gaussiana	1475,21	-715,60	826	22	**613,22**	**-284,61**	309	22

Analisando os resultados obtidos na estimação dos coeficientes dos modelos saturados para as duas sub-bases de dados (Tabela 4.13), concluí-se que quando se consideram todas as observações a progressão média do marcador varia ao longo do tempo havendo um incremento de 0,216 por cada ano. Em ambas os modelos verifica-se que pacientes com invasão vascular venosa apresentam valores iniciais da progressão mais elevados.

Tabela 4.13: Parâmetros estimados do modelo saturado com estrutura de correlação exponencial CA15-3: RT e RR

Estrutura de correlação exponencial	Base de dados RT		Sub-base de dados RR	
	β	p-valor	β	p-valor
Intercept	1,729	0,196	**2,225**	**0,032**
Tempo desde diagnóstico (anos)	**0,216**	**<0,0001**	0,055	0,181
Invasão vascular venosa (sim)	**1,200**	**0,034**	**1,493**	**0,001**
Idade ao diagnóstico (continua)	-0,007	0,723	-0,003	0,839
Bilateral (sim)	-1,010	0,091	-0,867	0,066
Estadio (III ou IV)	1,198	0,118	0,236	0,690
Grau (G3)	0,442	0,297	-0,032	0,922
Menopausa (sim)	0,290	0,631	0,256	0,582
Tipo de recidiva (metastizada)	-0,124	0,810	-0,025	0,950
Tipo de recidiva (local e metastizada)	1,064	0,260	1,122	0,124
Mama (esquerda)	0,018	0,955	-0,136	0,574
Presença de carcinoma associado (sim)	0,467	0,234	0,175	0,561
Invasão vascular linfática (sim)	0,006	0,988	-0,380	0,200
Recetor de estrogénio (positivo)	0,027	0,973	0,168	0,775
Recetor de progesterone (positivo)	0,327	0,513	0,496	0,206
Triplo negativo (sim)	0,014	0,987	0,677	0,311
Tumor primário (T3 ou T4 ou Tx)	-0,156	0,720	0,292	0,382
Nódulos linfáticos (Nx ou N0 ou N1)	0,667	0,377	0,257	0,656

Em conclusão, a presença de carcinoma associado e a idade ao diagnóstico parecem

influenciar as progressões quando se analisam todas as pacientes não sendo detetado o seu efeito quando se analisam apenas as pacientes com recidiva da doença.

4.2.2.1 Modelo final com estrutura de correlação exponencial para o marcador tumoral CA15-3

Considerando a base de dados total e a estrutura de correlação exponencial obteve-se o modelo apresentado na Tabela 4.14.

Tabela 4.14: Modelo final CA15-3 com estrutura de correlação exponencial

Estrutura de correlação exponencial	**Base de dados TT**	
	$\boldsymbol{\beta}$	**p-valor**
Intercept	2,310	<0,0001
Tempo desde diagnóstico (anos)	-0,003	0,731
Tempo desde diagnóstico:Recidiva	0,155	<0,0001
Idade ao diagnóstico (continua)	0,005	0,011
Recidiva (sim)	0,407	<0,0001
Invasão vascular venosa (sim)	0,657	<0,0001
$\boldsymbol{\hat{\nu}^2}$	0,352	
$\boldsymbol{\hat{\sigma}^2}$	0,354	
$\boldsymbol{\hat{\phi}}$	0,787	
$\boldsymbol{\hat{\tau}^2}$	0,031	
AIC	8406,235	
Log Likelohood	-4193,118	

Ao analisar-se os resultados obtidos concluí-se que pacientes sem recidivas não parecem apresentar variação dos valores do marcador ao longo do tempo ao contrário de pessoas com recidiva que para além de terem pontos iniciais mais elevados, cerca de 0,407, apresentam uma progressão do marcador mais acelerada aumentando 0,155 por cada ano. O valor 0,005, para idade ao diagnóstico, significa que há um aumento desse valor no log(CA15-3) por ano de idade no diagnóstico. Há um incremento de 0,657 no ponto inicial da progressão média dos valores do logaritmo do marcador tumoral em pacientes com invasão linfática venosa. A estrutura de correlação que melhor representa a variabilidade dos dados é aquela que incorpora efeitos aleatórios ao nível individual com $\hat{\nu}^2 = 0,352$, uma estrutura de correlação exponencial para descrever a variabilidade dentro dos pacientes com $\hat{\rho}^2 = \exp(-\frac{1}{0,787}|u|)$ e $\hat{\sigma}^2 = 0,354$ e um erro de medição com variância $\hat{\tau}^2 = 0,031$.

4.2.2.2 Análise de diagnóstico

Para a verificação dos pressupostos do modelo analisam-se as Figuras 4.28 e 4.29 que apresentam o gráfico dos resíduos específicos do sujeito *versus* os valores ajustados e um gráfico Q-Q normal dos resíduos específicos do sujeito, respetivamente. Observando-se estes gráficos, verifica-se que na Figura 4.29 há um desvio dos valores em relação á reta, no entanto, os pressupostos do modelo longitudinal, relativos à homogeneidade das

variâncias dos erros de medição e à normalidade destes, não parecem passíveis de rejeição. A adequabilidade da estrutura de correlação escolhida já foi analisada anteriormente.

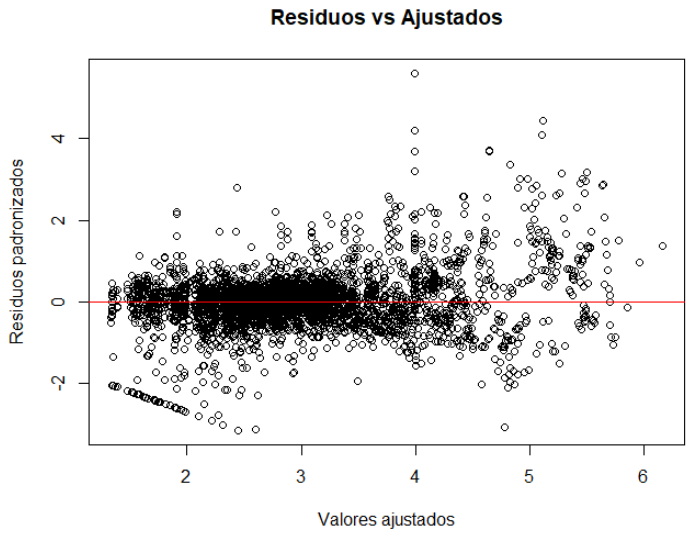

Figura 4.28: Resíduos específicos do sujeito *versus* os valores ajustados: CA15-3

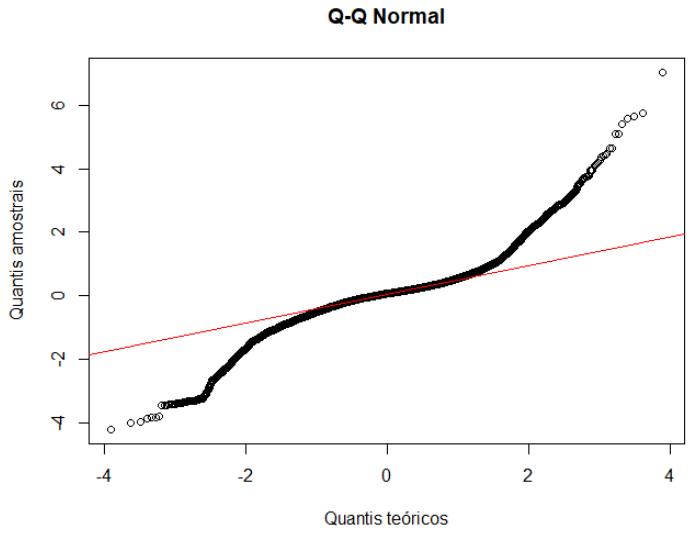

Figura 4.29: Q-Q dos resíduos específicos do sujeito: CA15-3

4.3 Análise de sobrevivência

Para a análise de sobrevivência considera-se como evento de interesse a recidiva do cancro da mama. O tempo até ao evento de interesse será a diferença entre a data de diagnóstico e a data da recidiva. Em alguns casos, a data de diagnóstico não está disponível, nessas 7 pacientes utilizou-se a data da primeira consulta para o cálculo do tempo até ao evento de interesse. Como se pode observar na Figura 4.30, onde é apresentada a diferença entre a data de diagnóstico e a data da primeira consulta, esta última parece ser uma boa aproximação quando não há informação sobre a data de diagnóstico do cancro da mama da paciente, uma vez que das pacientes que tem disponíveis as duas datas, 97,5% das diferenças é zero, significando que as datas coincidem. As pacientes que apresentam diferenças positivas são pacientes que começaram a ser seguidas antes do diagnóstico, por haver suspeita de um possível cancro no futuro, e as pacientes que tem diferenças negativas tiveram a primeira consulta no Hospital de Braga após o diagnóstico de cancro da mama.

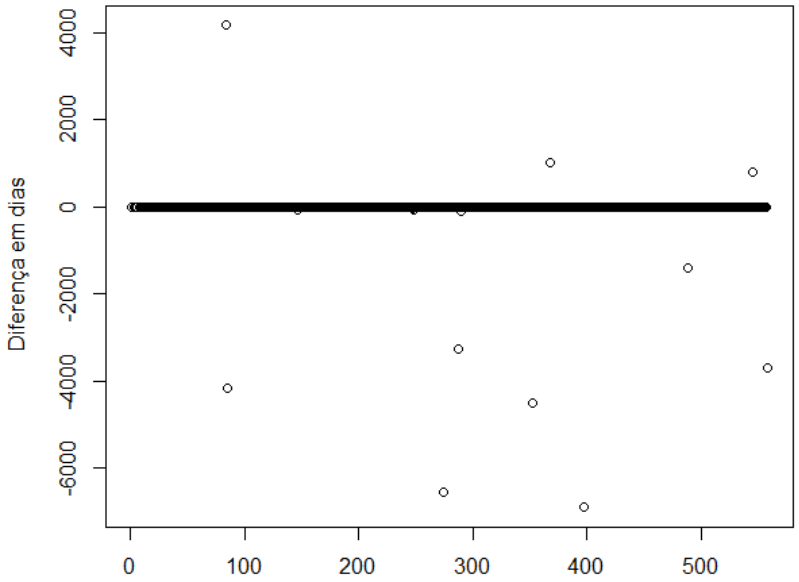

Figura 4.30: Diferença entre a data de diagnóstico e a data da primeira consulta

Como já foi referido anteriormente, a Unidade de Senologia do Hospital de Braga foi criada em 2008. Todas as pacientes que estavam a ser seguidas no hospital à data de 1 de janeiro de 2008 fazem parte desta amostra. Assim, será importante considerar-se truncatura à esquerda, uma vez que se perdeu a informação das pacientes que morreram antes do início da unidade e, asim, apenas fica disponível a informação das mulheres que tiveram uma melhor resposta ao tratamento. A Figura 4.31 representa as curvas

de Kaplan-Meier, considerando como evento de interesse a recidiva da doença, tendo em conta apenas censura pela direita e tendo em conta censura pela direita e truncatura pela esquerda. Para esta análise de sobrevivência foram consideradas as 559 pacientes que constituem a base de dados inicial mais completa.

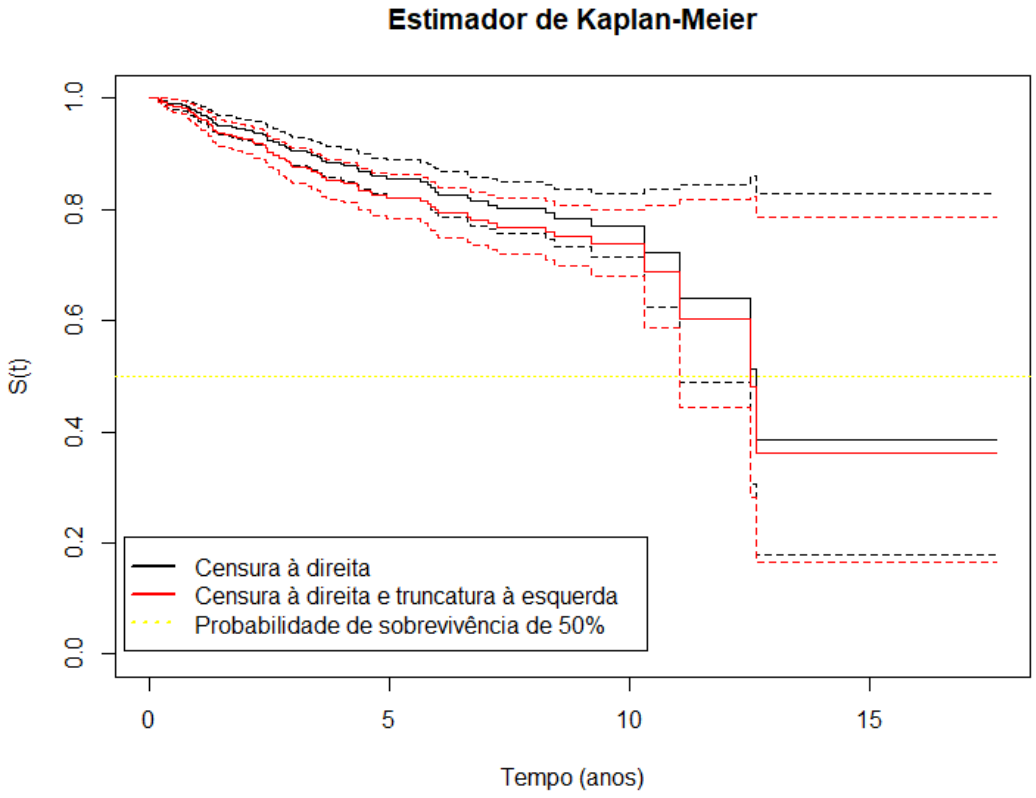

Figura 4.31: Curvas de Kaplan-Meier considerando censura pela direita e considerando censura pela direita e truncatura pela esquerda.

As curvas apresentadas mostram as probabilidades de sobrevivência estimadas contra o tempo em anos. Esta função é uma função em escada que toma o valor 1 no tempo 0 e decresce apenas com os eventos. As observações censuradas no estimador de Kaplan-Meier têm peso nulo e são estas observações que fazem com que a função não alcance o valor zero. As linhas a tracejado representam os intervalos de confiança associados às diferentes curvas de sobrevivência. O tempo de vida mediano que representa o tempo em que a probabilidade de sobrevivência é 0,5 é no caso de se considerar apenas censura pela direita é 12,6 e 12,5 se se considerar truncatura pela esquerda e censura pela direita. Observando as curvas, percebe-se que há uma sobrestimação das probabilidades de sobrevivência quando se ignora a truncatura à esquerda. Isto é explicado pelo facto de apenas as pacientes com melhor resposta ao tratamento estarem a contribuir para a estimação da curva de sobrevivência e, ignorando a truncatura, todas tem a mesma contribuição. Uma vez que um dos principais interesses deste projeto é concluir sobre a adequabilidade

da utilização dos modelos conjuntos e, até ao momento, não sendo possível incorporar mecanismos de truncatura nos *packages* que serão utilizar para este efeito, optou-se por seguir as análises apenas considerando a censura pela direita. Assim, deve-se sempre ter em consideração que há uma ligeira sobrestimação das probabilidades de sobrevivência.

As covariáveis discretas podem ser incluídas no estimador de Kaplan-Meier, assim, para cada nível da covariável é aplicado o estimador produto-limite que calcula a probabilidade de observar o evento de interesse num dado instante de tempo, condicionado à sobrevivência até aquele ponto.

Tabela 4.15: Teste igualdade de curvas

Variável	**p-valor**
Menopausa	0,030
Recetor de estrogénio	<0,0001
Recetor de progesterone	0,001
Triplo negativo	<0,0001
Estadio	<0,0001
Invasão vascular linfática	0,010
Invasão vascular venosa	0,001
Grau Bloom e Richardson	<0,0001
Tumor primário	<0,0001
Grau de disseminação	<0,0001
Terapia hormonal	<0,0001
Tratamento primário	<0,0001
Tratamento cirúrgico	<0,0001
Tipo cirurgia	0,001

A Tabela 4.15 apresenta o resultado do teste sobre a igualdade das curvas de sobrevivência para os diferentes níveis de cada variável. As 14 variáveis apresentadas são as que revelaram ter diferenças significativas nas curvas de sobrevivência, sob a hipótese nula de que as diferentes categorias apresentam curvas de sobrevivência iguais. Nas variáveis estadio, invasão vascular venosa, grau de disseminação, tratamento cirúrgico e tratamento primário foi aplicado o teste de Log-Rank e nas demais o teste de Gehan-Wilcoxon com modificação de Peto e Peto, visto que o pressuposto de riscos proporcionais necessário para a aplicação do teste de Log-Rank não era verificado.

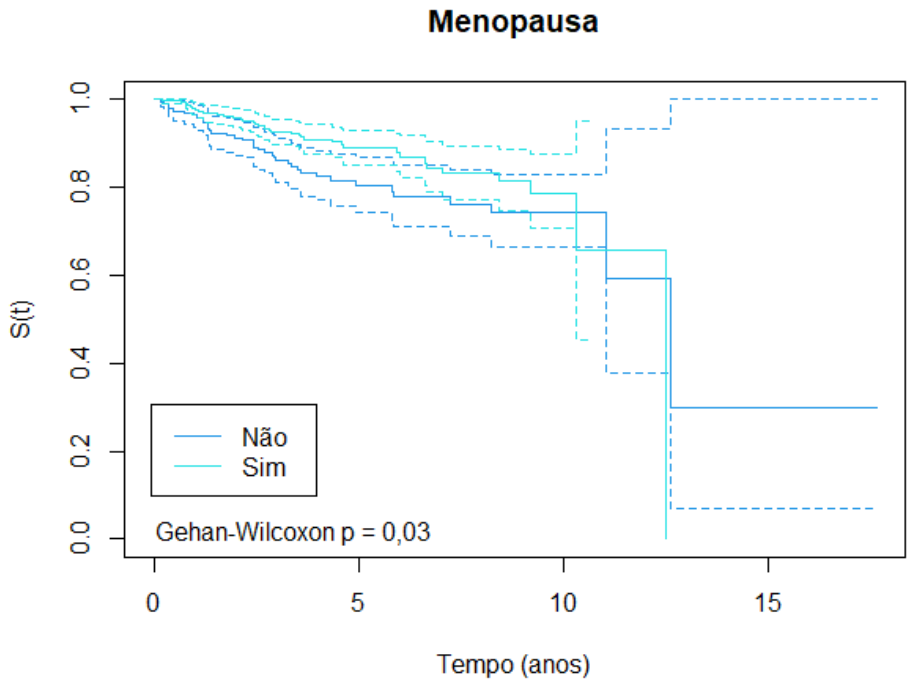

Figura 4.32: Curvas de Kaplan-Meier: Menopausa

A Figura 4.32 apresenta as curvas de sobrevivência para pacientes com e sem menopausa. O teste de Gehan-Wilcoxon aplicado comprova que as probabilidades de recidiva da doença são diferentes entre os dois grupos de pacientes e consegue-se identificar através da análise gráfica que mulheres que estão na menopausa apresentam probabilidades de recidiva menores. Assim, qualquer que seja o instante t, a probabilidade de não recidivar pelo menos até esse instante é superior nas mulheres que estão na menopausa.

Através da análise da Figura 4.33 concluí-se que pacientes com recetores de estrogénio positivo tem probabilidades de recidiva menores em comparação com pacientes que apresentam recetores de estrogénio negativos. Em relação aos recetores de progesterona (Figura 4.34), pacientes que testam positivo tem probabilidade de recidivar a doença mais baixa que pacientes com recetores de progesterona negativos.

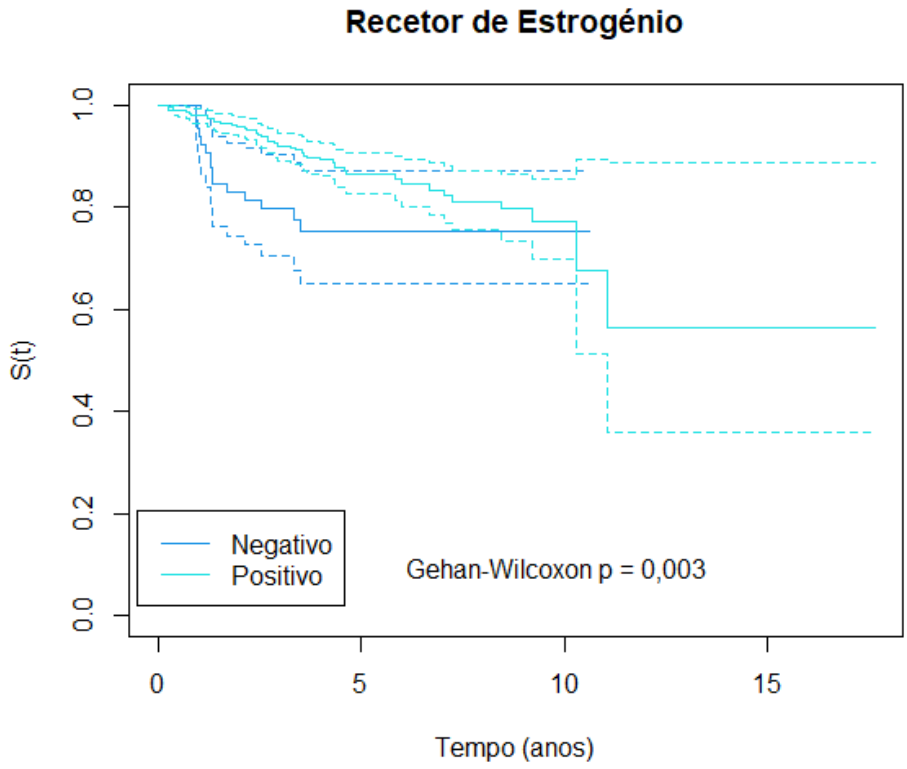

Figura 4.33: Curvas de Kaplan-Meier: Recetores de estrogénio

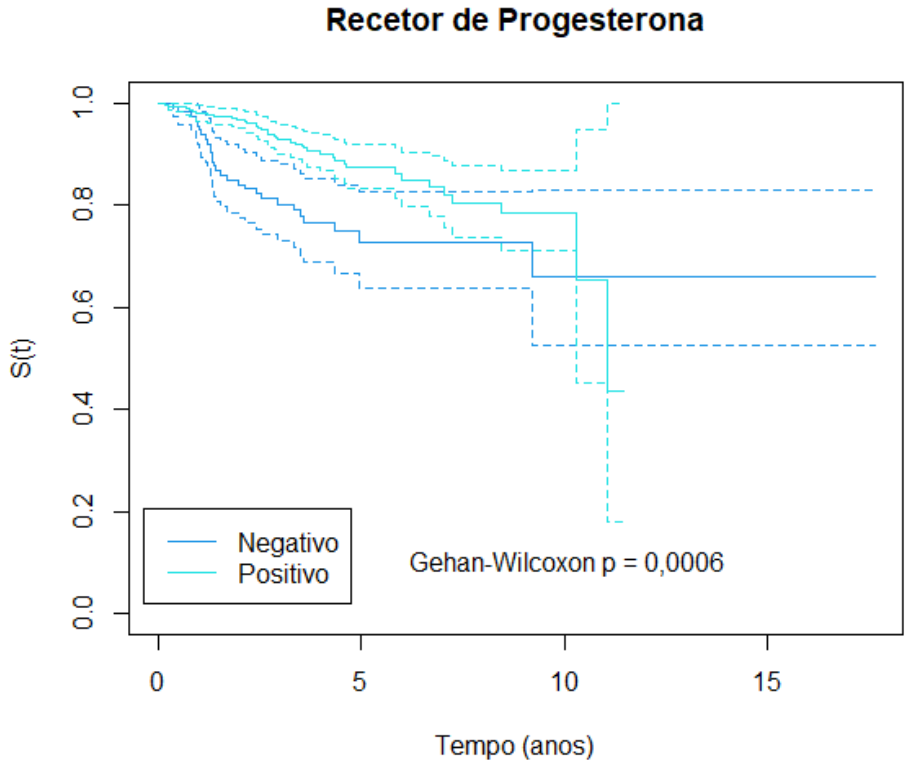

Figura 4.34: Curvas de Kaplan-Meier: Recetores de progesterona

Pacientes com subtipo triplo negativo apresentam maiores probabilidades de recidivar a doença. A Figura 4.35 apresenta a estimação das curvas de sobrevivência para pacientes com ou sem este subtipo de cancro. A grande amplitude dos intervalos de confiança no final das curvas deve-se ao baixo número de pacientes considerados para a sua estimação.

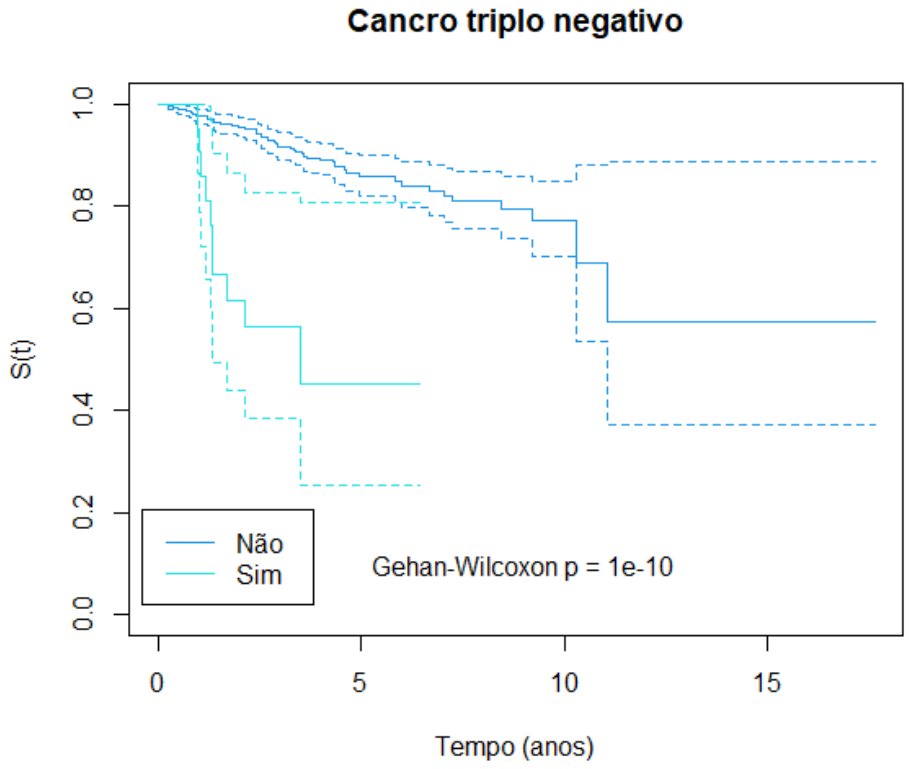

Figura 4.35: Curvas de Kaplan-Meier: Triplo negativo

Como já foi referido na análise exploratória, o estadio é uma combinação das categorias T (extensão do tamanho do tumor), N (estado nodal) e M (presença ou ausência de metástases à distância). Inicialmente, estavam definidas 5 categorias cujas curvas de sobrevivência estimadas são apresentadas na primeira imagem da Figura 4.36. Contudo, através da realização de comparações múltiplas usando o método de Benjamini e Hochberg (1995), conseguiu-se identificar que as curvas dos estadios 0, I e II podem ser consideradas iguais e as curvas dos estádios III e VI também não apresentam diferenças estatísticamente significativas. Assim, construiram-se novas categorias. A Figura 4.37 apresenta as curvas de sobrevivência estimadas para as categorias 0_I_II e III_IV. Através da análise destas curvas identifica-se que pacientes com estadio 0_I_II tem probabilidades de recidiva da doença inferiores.

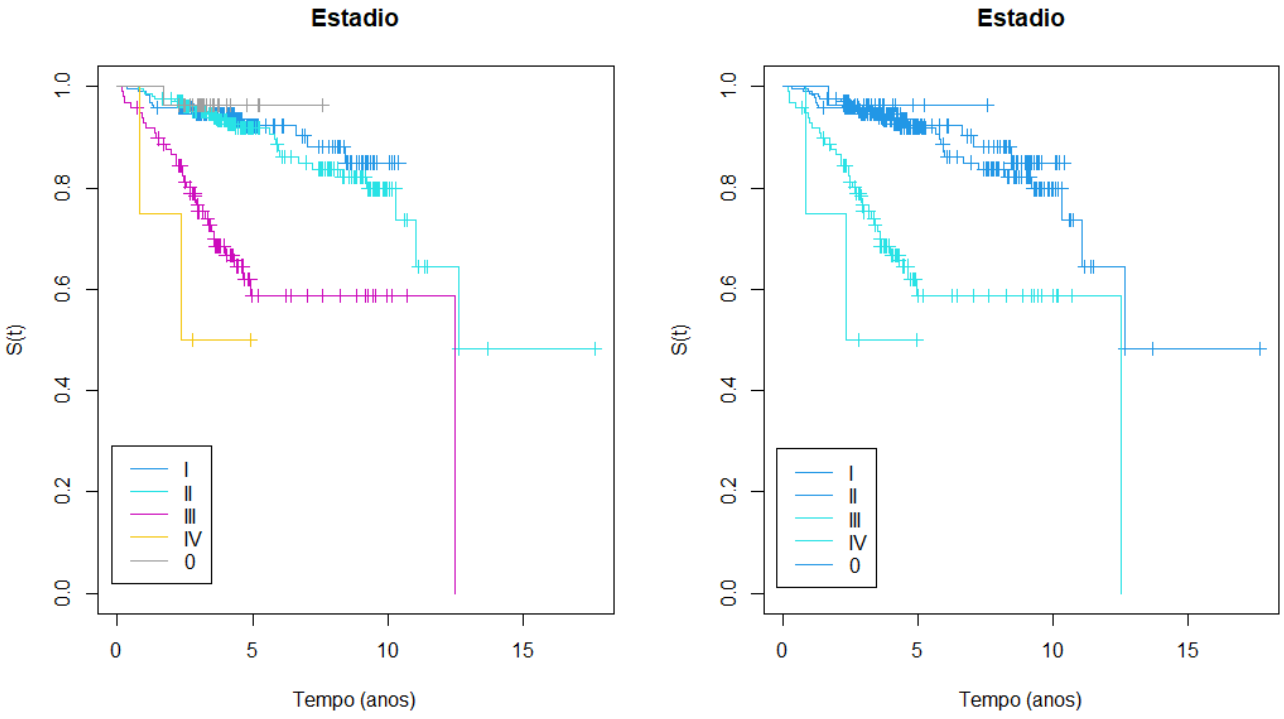

Figura 4.36: Curvas de Kaplan-Meier: Identificar grupos estadio

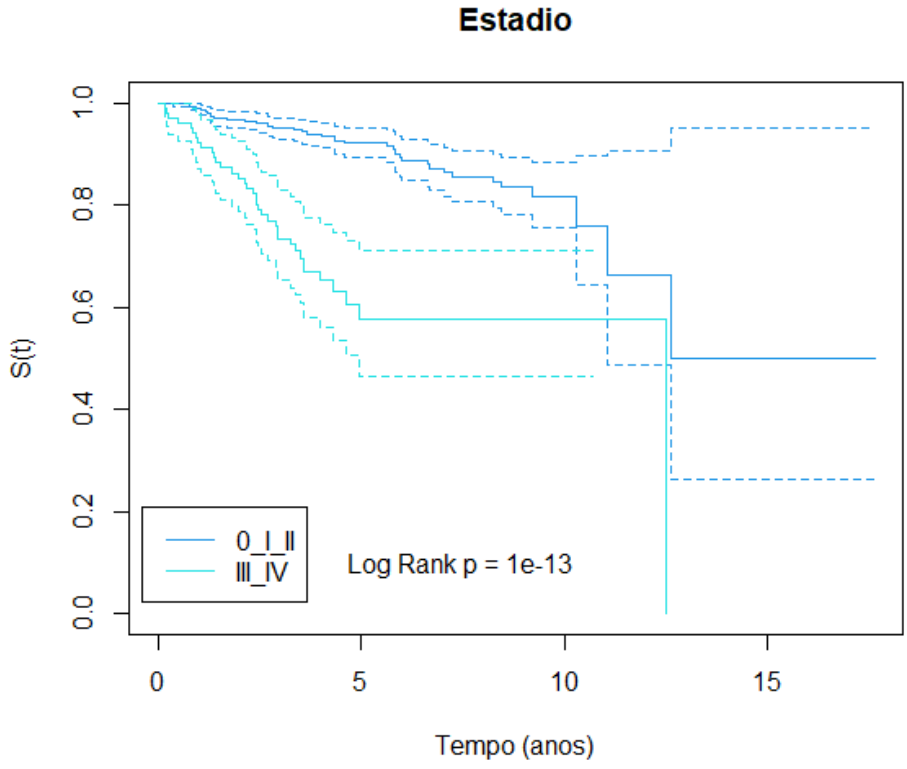

Figura 4.37: Curvas de Kaplan-Meier: Estadio

A presença de invasão linfática traduz-se numa probabilidade de recidiva do cancro da mama maior (Figura 4.38), assim como a presença de invasão venosa (Figura 4.39). A diferença das curvas de sobrevivência para as duas categorias em ambas as variáveis é comprovada pelo p-valor apresentado quer na Tabela 4.15 quer nas Figuras apresentadas.

O grau de Bloom e Richardson apresenta 4 categorias diferentes: Grau I (baixa malignidade); Grau II (malignidade intermédia), Grau III (alta malignidade) e Gx (não foi possível identificar o grau do tumor). A Figura 4.40 apresenta as diferentes curvas de sobrevivência estimadas para as diferentes categorias. Após a realização de comparações múltiplas usando o método de Benjamini e Hochberg (1995), conseguiu-se criar dois grupos distintos visto que as curvas de sobrevivência dos graus G1, G2 e Gx podem ser consideradas iguais. Pacientes com grau III, ou seja, com alta malignidade do tumor e, na prática, com maior probabilidade de desenvolver metástases apresentam probabilidades de recidiva mais elevadas (Figura 4.41).

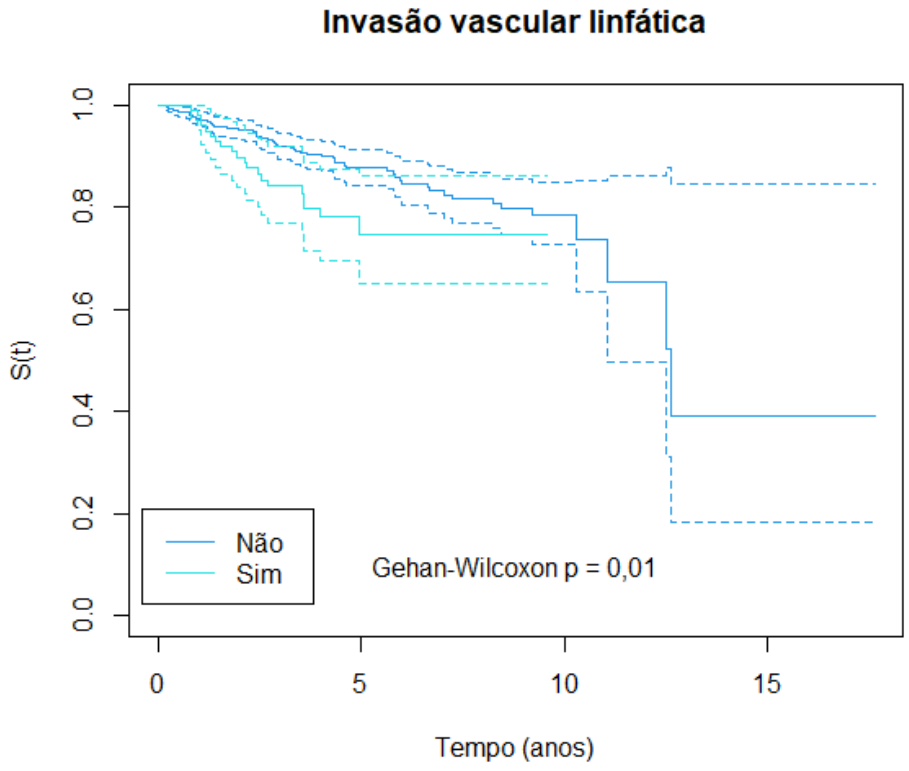

Figura 4.38: Curvas de Kaplan-Meier: Invasão vascular linfática

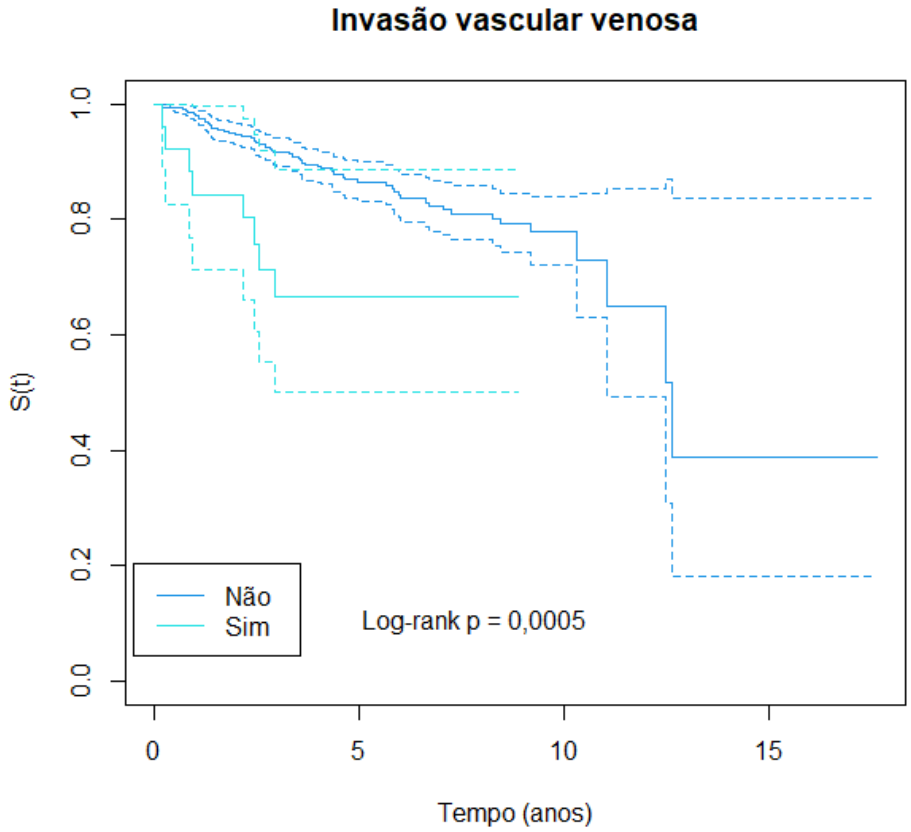

Figura 4.39: Curvas de Kaplan-Meier: Invasão vascular venosa

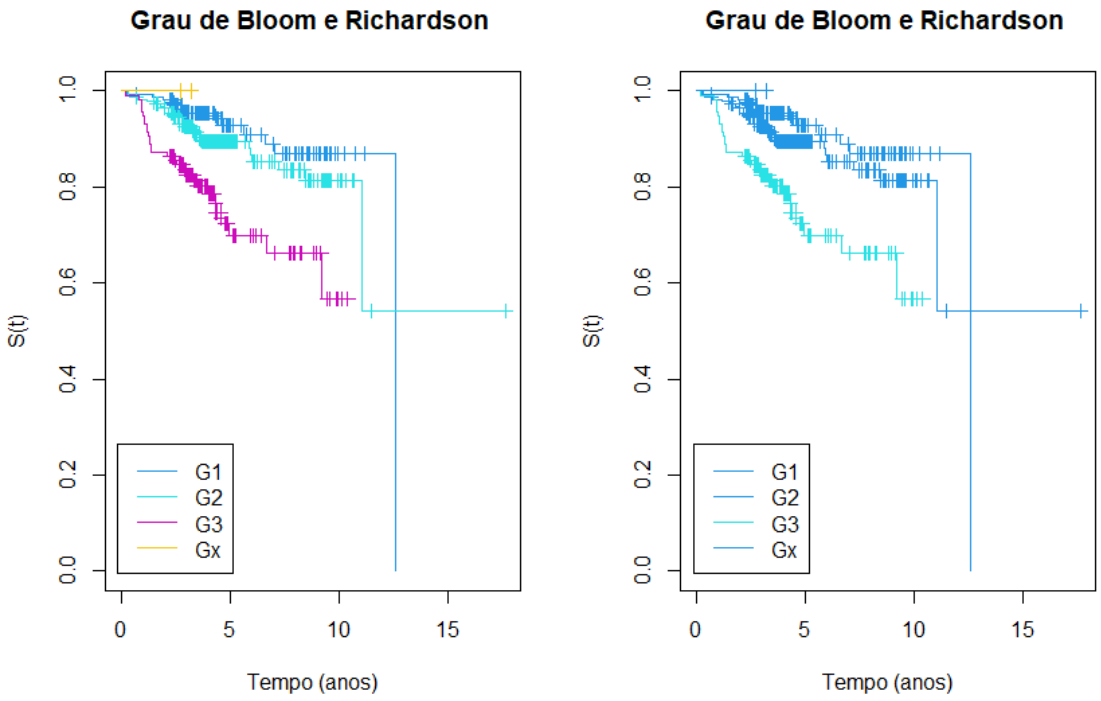

Figura 4.40: Curvas de Kaplan-Meier: Identificar grupos grau de Bloom e Richardson

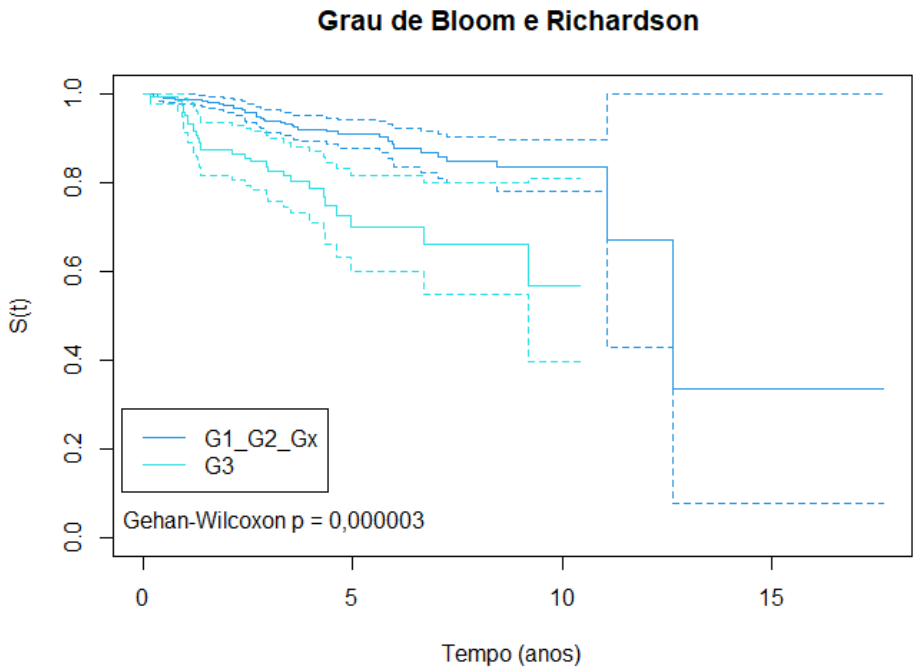

Figura 4.41: Curvas de Kaplan-Meier: Grau de Bloom e Richardson

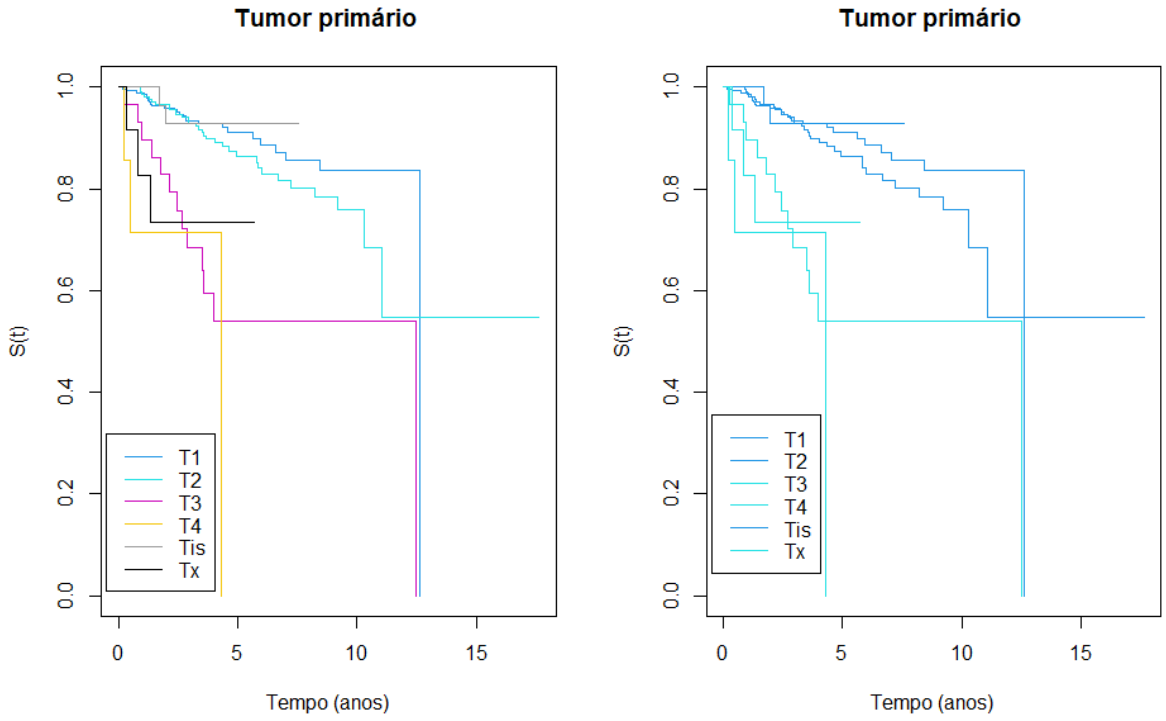

Figura 4.42: Curvas de Kaplan-Meier: Identificar grupos tumor primário

A Figura 4.42 apresenta as curvas de sobrevivência referentes as diferentes categorias do tumor primário. Após a realização de comparações múltiplas usando o método de

Benjamini e Hochberg (1995), criaram-se duas novas categorias T1_T2_Tis e T3_T4_Tx visto que não há diferença significativas entre as curvas de sobrevivência das categorias T1, T2 e Tx e entre as curvas de sobrevivência das categorias T3, T4 e Tis. A Figura 4.43 mostra que pacientes cujo tumor primário é classificado como T3_T4_Tx apresentam probabilidades de recidiva do cancro da mama mais elevadas.

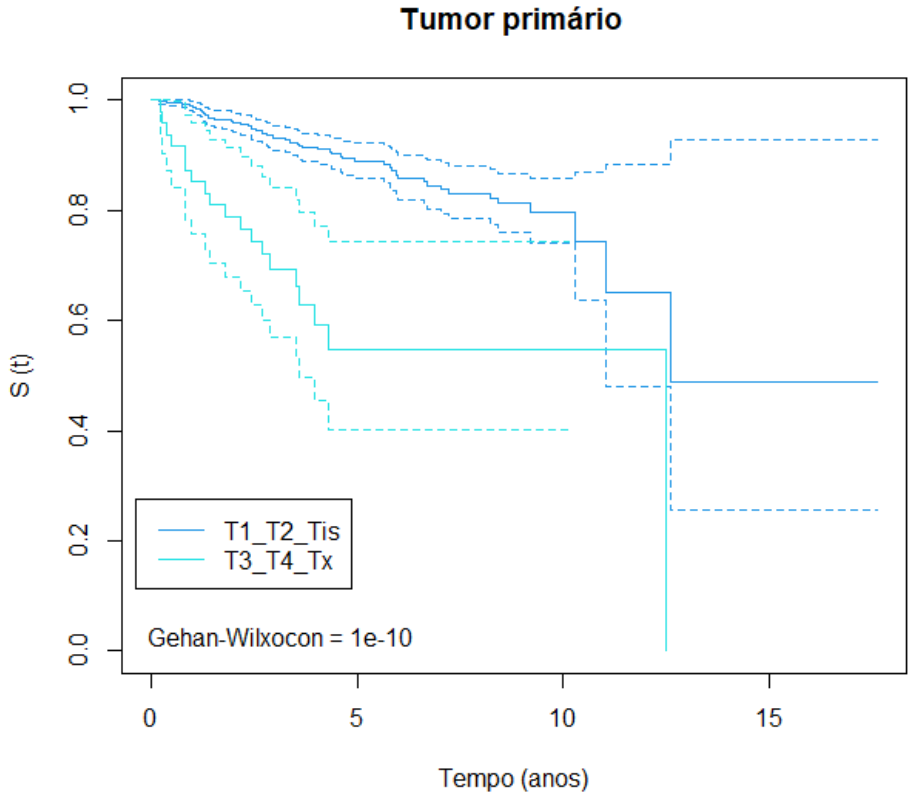

Figura 4.43: Curvas de Kaplan-Meier:Tumor primário

As quatro categorias do grau de disseminação dos nódulos linfáticos foram analisadas e após a realização de comparações múltiplas usando o método de Benjamini e Hochberg (1995), criaram-se duas novas categorias. A Figura 4.44 apresenta as curvas de sobrevivência para todas as categorias iniciais e a Figura 4.45 as curvas de sobrevivência para as novas categorias. Concluí-se através da análise da Figura 4.45 que pacientes com grau de disseminação N2 ou N3 têm probabilidades de recidivar a doença mais elevadas.

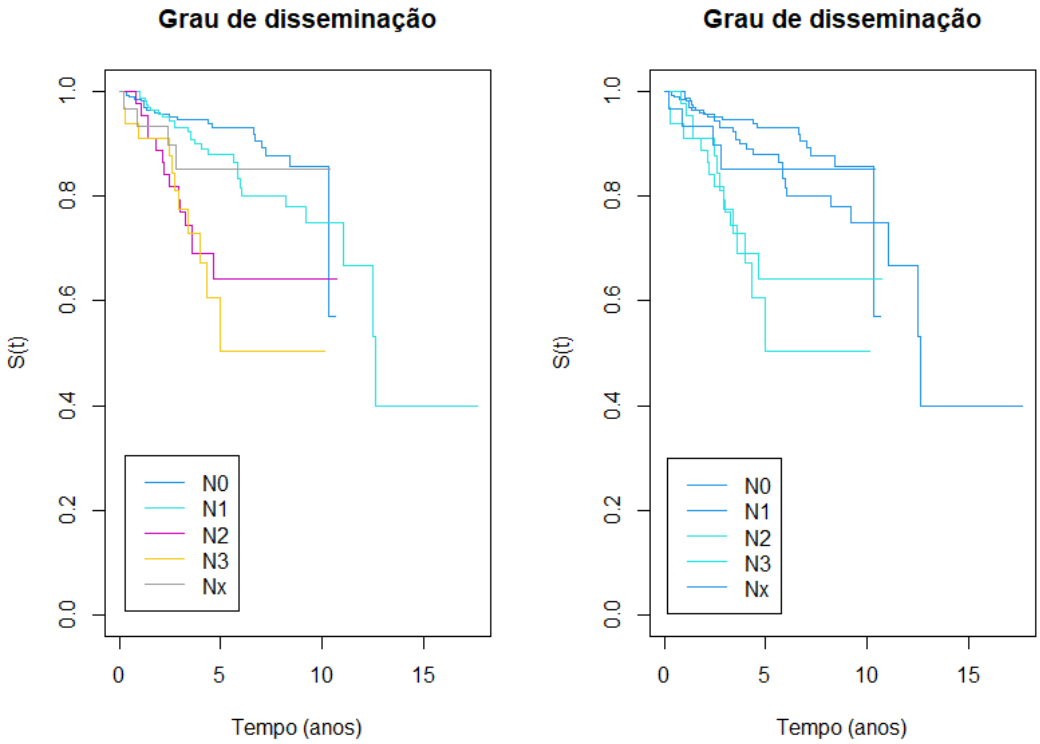

Figura 4.44: Curvas de Kaplan-Meier: Identificar grupos grau de disseminação

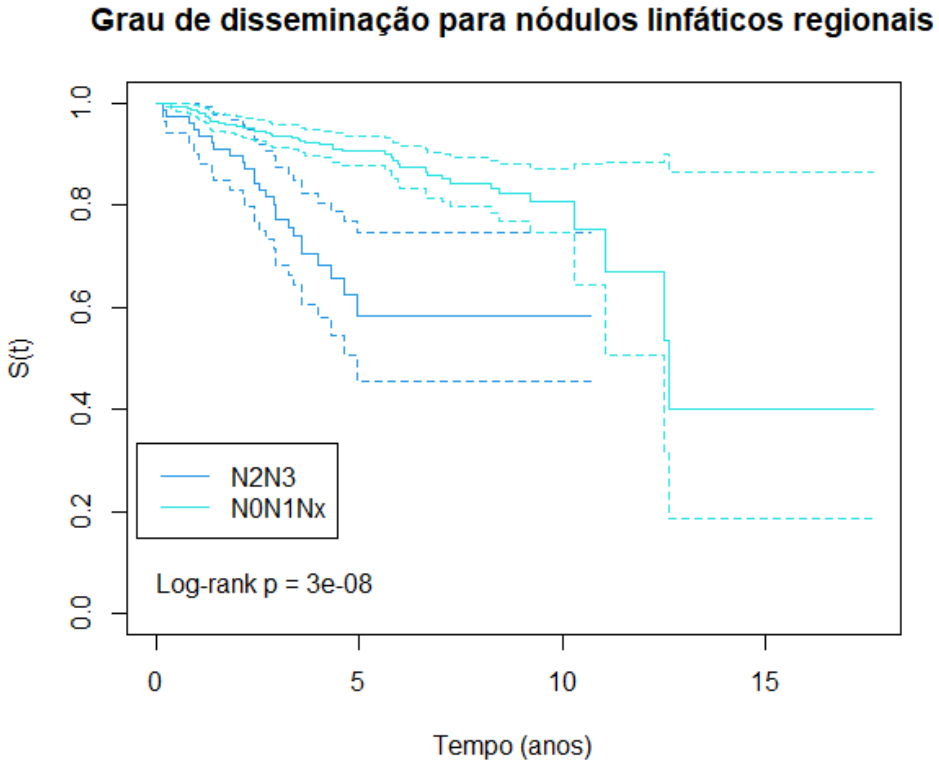

Figura 4.45: Curvas de Kaplan-Meier: Grau de disseminação

Em relação às variáveis associadas ao tratamento aplicado, a Figura 4.46 apresenta as curvas de sobrevivência para as pacientes que tiveram terapia hormonal e para as que não tiveram. As mulheres que não tiveram terapia hormonal apresentam probabilidades de recidiva mais elevadas. Como se pode verificar na Figura 4.47 pacientes que foram submetidas a tratamento primário e pacientes que não foram têm probabilidades de recidiva do cancro da mama bastante diferentes. As curvas apresentadas mostram que pacientes com tratamento primário tem maior probabilidade de recidiva da doença. Importante referir que isto não será um resultado da aplicação do tratamento, mas sim do estado inicial da paciente que leva à aplicação deste tratamento.

Relativamente ao tratamento cirúrgico, a Figura 4.48 apresenta as curvas de sobrevivência para pacientes que foram intervencionadas cirurgicamente e para pacientes que não foram. As pacientes que foram intervencionadas tem probabilidades de recidiva inferiores. A fim de se concluir sobre a probabilidade de recidiva com base nos dois tipos de cirurgia, apresenta-se a Figura 4.49 que representa as curvas de sobrevivência para as pacientes submetidas a uma cirurgia conservadora da mama e pacientes que fizeram mastectomia com ou sem dissecção dos linfonodos axilares. Pacientes submetidas a uma mastectomia tem probabilidades de recidiva mais elevadas, mais uma vez, a sobrevivência mais baixa não significa que a cirurgia seja a causa da diminuição da sobrevivência. Significa, sim, que casos mais graves de cancro necessitam de cirurgia e por serem mais graves tem maior probabilidade de recidiva.

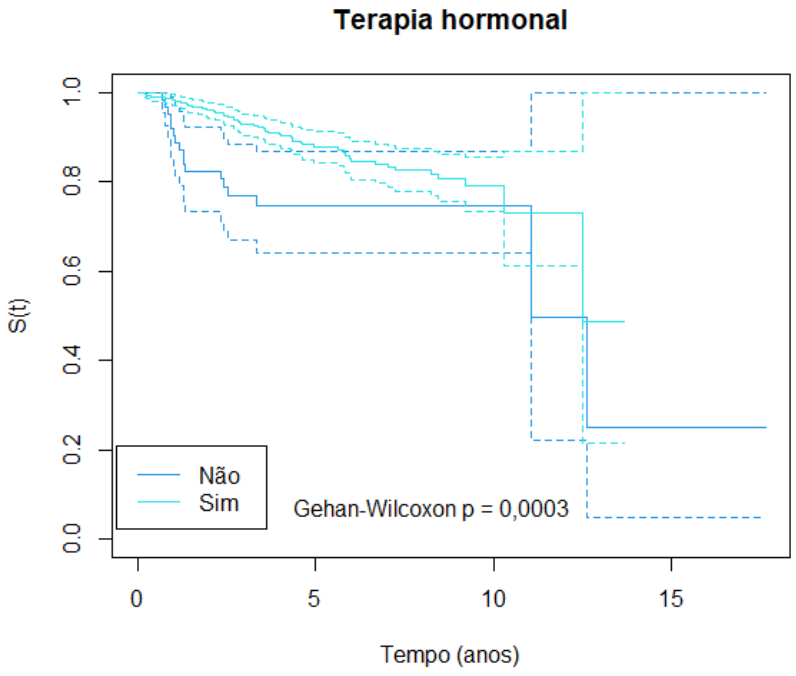

Figura 4.46: Curvas de Kaplan-Meier: Terapia hormonal

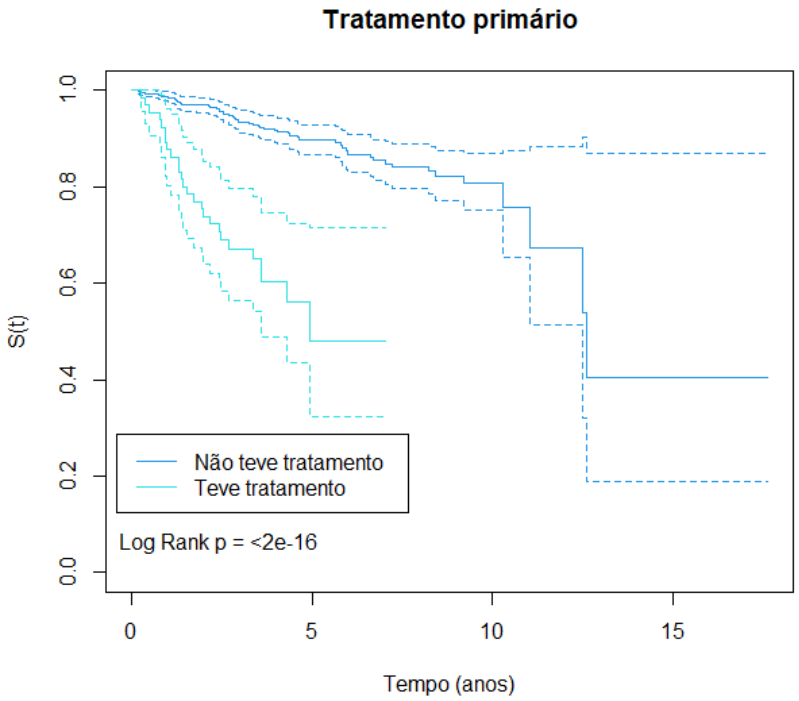

Figura 4.47: Curvas de Kaplan-Meier: Tratamento primário

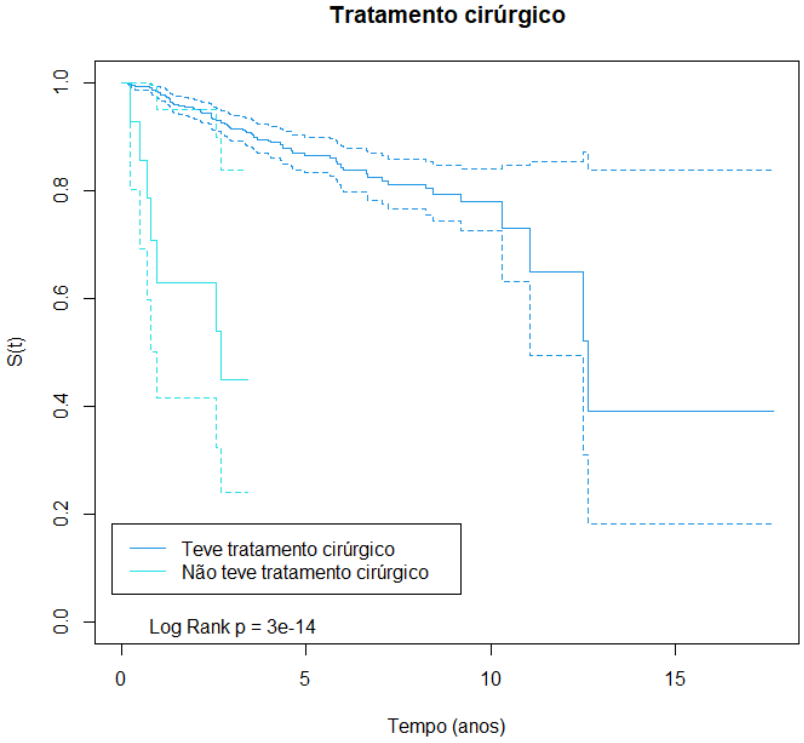

Figura 4.48: Curvas de Kaplan-Meier: Tratamento cirúrgico

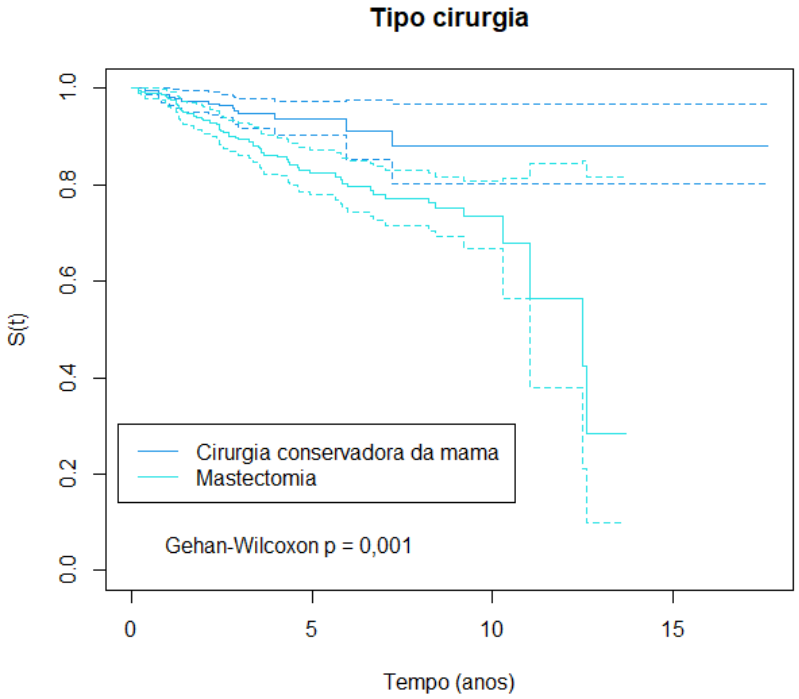

Figura 4.49: Curvas de Kaplan-Meier: Tipo de cirurgia

4.3.1 Modelo de regressão de Cox

O modelo de regressão de Cox (1972) permite comparar grupos de indivíduos com valores de covariáveis diferentes. A interpretação dos valores da *hazard ratio* (HR) ($\exp(\beta)$) permite comparar o risco de diferentes grupos de indivíduos. A Tabela 4.16 apresenta os resultados das estimações da HR para modelo de regressão simples das variáveis que mostraram ter riscos de recidiva diferentes para os diferentes níveis.

Ao nível do tratamento, mulheres intervencionadas cirurgicamente tem 92% menor risco de ter recidiva da doença comparativamente com mulheres que não foram submetidas a nenhuma cirurgia. Ainda se consegue concluir que uma mulher submetida a uma mastectomia tem cerca de 2,6 vezes maior risco de recidivar a doença que uma mulher que fez cirurgia conservadora da mama. Mas, uma mulher que não tenha sido intervencionada cirurgicamente tem cerca de 20 vezes maior risco de recidivar a doença que uma mulher que fez cirurgia conservadora da mama. Pacientes que tiveram tratamento hormonal apresentam cerca 60% menor risco de recidivar a doença comparativamente com mulheres que não fazem este tipo de tratamento. Pacientes que foram submetidas a tratamento primário tem cerca de 6 vezes maior risco de recidiva do cancro da mama comparativamente com pacientes que não foram submetidas. A menopausa, recetores de estrogénio positivos, recetores de progesterona positivos e nódulos linfáticos Nx ou N0 ou N1, apresentam um efeito protetor. O aumento da idade ao diagnóstico representa, também, um efeito protetor, assim por cada ano a mais na idade de diagnóstico, o risco de

apresentar recidiva da doença diminui cerca de 2%. O subtipo triplo negativo, estadio da doença mais avançado (III ou VI), invasão vascular linfática ou venosa, tumor primário não encontrado, de grande dimensão ou de qualquer tamanho com extensão direta para a parede torácica ou pele e alta malignidade no grau de Bloom e Richardson apresentam um risco mais elevado de desenvolver uma recidiva do cancro da mama.

Tabela 4.16: Modelo de regressão de Cox simples

Variável	HR	p-valor
Menopausa (sim)	0,636	0,048
Recetor de estrogénio (positivo)	0,442	0,006
Recetor de progesterona (positivo)	0,461	0,002
Triplo negativo (sim)	6,763	<0,0001
Estadio (III ou IV)	4,659	<0,0001
Invasão vascular linfática (sim)	1,920	0,013
Invasão vascular venosa (sim)	3,422	0,001
Tumor primário (T3 ou T4 ou Tx)	4,670	<0,0001
Nódulos linfáticos (Nx ou N0 ou N1)	0,280	<0,0001
Grau Bloom e Richardson (G3)	2,983	<0,0001
Idade ao diagnóstico (continua)	0,979	0,014
Terapia hormonal (sim)	0,392	0,001
Tratamento primário (sim)	6,119	<0,0001
Cirurgia (sim)	0,097	<0,0001
Tipo de cirurgia (mastectomia)	2,578	0,001
Tipo de cirurgia (sem cirurgia)	19,843	<0,0001

Partindo-se de um modelo de Cox incluindo todas as variáveis apresentadas na tabela anterior à exceção das variáveis referentes a tratamento, encontrou-se o modelo de regressão múltipla apresentado na Tabela 4.17.

Tabela 4.17: Modelo de regressão de Cox múltiplo

Variáveis	HR	p-valor
Triplo negativo (sim)	4,639	<0,0001
Estadio (III ou IV)	4,279	<0,0001
Grau (G3)	1,996	0,018
Idade ao diagnóstico (continua)	0,974	0,007
AIC	609,520	

Do modelo de regressão múltipla apresentado consegui-se concluir que pacientes com subtipo triplo negativo apresentam 4,6 vezes maior risco de ter recidiva do cancro da mama comparativamente com mulheres que não apresentam este subtipo molecular. Mulheres com estadio da doença classificado com estadio III ou IV tem 4,3 vezes maior risco de recidivar a doença que pacientes com mulheres com estadio 0 ou I ou II. Mulheres com Grau III na classificação de Bloom e Richardson (alta malignidade), apresentam cerca de 2 vezes maior risco de recidivar a doença comparativamente com mulheres com grau I, II ou Gx. Por cada ano a mais na idade de diagnóstico, o risco de apresentar recidiva da doença diminui cerca de 2,6%.

4.3.1.1 Análise de diagnóstico

A Figura 4.50 mostra que o pressuposto dos riscos proporcionais necessário para a correta aplicação do modelo de Cox é verificado, dado que se verifica uma nuvem de pontos aleatórios em cada uma das quatro representações gráficas apresentadas. Assim, concluí-se que o pressuposto dos riscos proporcionais é verificado, uma vez que não se identifica nenhum padrão entre os resíduos e o tempo. A Figura 4.51 representa os resíduos de Cox-Snell *versus* a estimativa de Nelson-Aalen da função de risco cumulativa dos resíduos. Como a maioria dos pontos sobrepõem a recta de declive um e ordenada na origem nula, concluí-se que o modelo obtido é adequado.

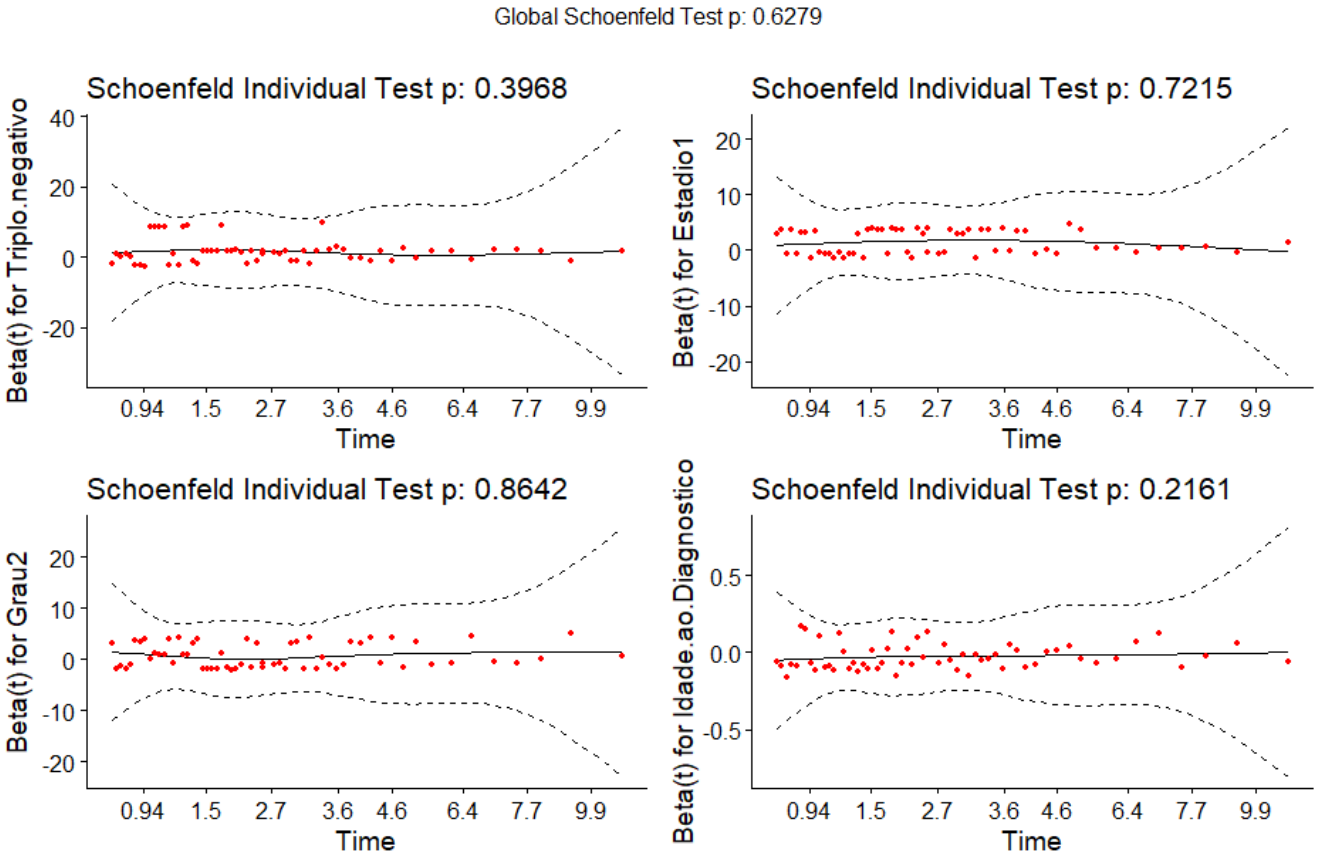

Figura 4.50: Resíduos de Schoenfeld

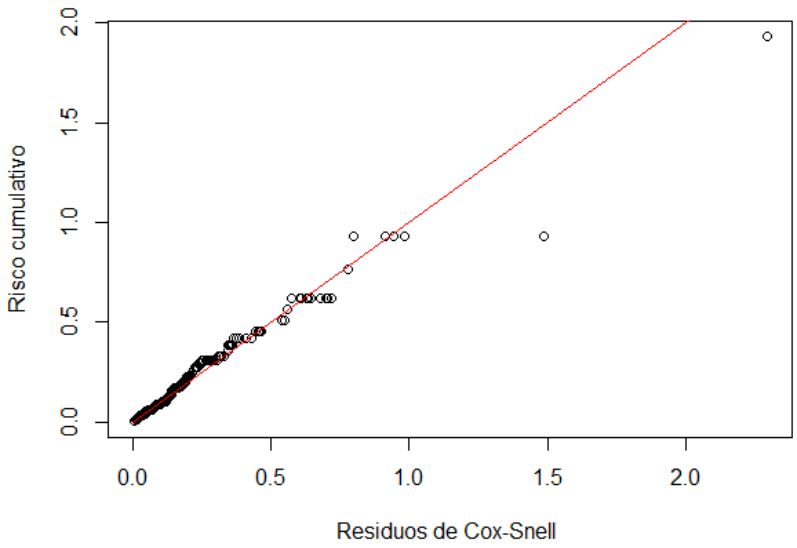

Figura 4.51: Resíduos Cox-Snell

4.4 Análise modelos conjuntos

4.4.1 Marcador tumoral CEA

O modelo conjunto do processo longitudinal dos valores do marcador tumoral CEA e o processo de sobrevivência, foi ajustado utilizando os *packages joineR* e *JM*, como referido no Capítulo 3.

Semelhantemente à análise longitudinal apresentada na secção 4.2.1, foi utilizada uma transformação logarítmica dos valores do marcador tumoral CEA no processo longitudinal em ambos os modelos ajustados. É importante referir que as bases de dados utilizadas para a modelação conjunta têm menos pacientes que as consideradas nas análises longitudinais e de sobrevivência efetuadas anteriormente em separado. Para o processo longitudinal apenas foram consideradas as observações até à recidiva. Assim, para as análises conjuntas consideraram-se 396 pacientes e um total de 2713 observações do marcador tumoral CEA, visto que foram retirados os valores faltantes da base de dados utilizada. Desta forma, uma comparação direta entre os resultados obtidos com os modelos conjuntos e com os modelos obtidos em separado para cada processo não é adequada visto que o número de pacientes considerados e as suas próprias características afetam a estimação dos parâmetros.

Como já foi referido, no processo de sobrevivência partiu-se com as variáveis que mostraram significância estatística na análise de sobrevivência efetuada anteriormente, e na parte longitudinal partiu-se de um modelo saturado (16 variáveis e uma interação utilizadas no processo longitudinal anterior) e foi-se retirando as variáveis que não mostravam significância estatística no processo longitudinal no modelo conjunto ajustado. O modelo

foi inicialmente ajustado com o *package JM* e após se encontar o modelo só com as variáveis significativas ajustou-se um modelo com o *package joineR*, para se compararem os resultados obtidos com os dois *packages*.

Tabela 4.18: Modelos conjuntos *packages* joineR e JM: CEA

Modelos conjuntos	**joineR**			**JM**	
(sub-base de dados TR)	**Estimativa**	**Inferior 95%**	**Superior 95%**	**Estimativa**	**p-valor**
Processo Longitudinal					
Intercept	-0,682	-1,645	0,044	**-1,052**	**<0,0001**
Tempo desde diagnóstico (anos)	**0,038**	**0,020**	**0,051**	**0,037**	**<0,0001**
Tempo desde diagnóstico:Recidiva	**0,179**	**0,108**	**0,247**	**0,213**	**<0,0001**
Idade ao diagnóstico (continua)	**0,017**	**0,010**	**0,025**	**0,016**	**<0,0001**
Estadio (III ou IV)	0,258	-0,317	0,904	**0,437**	**<0,0001**
Grau (G3)	-0,015	-0,182	0,110	**0,193**	**<0,0001**
Menopausa (sim)	-0,184	-0,423	0,019	**-0,105**	**0,035**
Recidiva (sim)	-0,169	-0,604	0,147	0,007	0,900
Mama (esquerda)	0,033	-0,113	0,131	**0,115**	**<0,0001**
Presença de carcinoma associado (sim)	0,081	-0,082	0,203	**0,128**	**<0,0001**
Invasão vascular linfática (sim)	0,025	-0,140	0,164	**0,059**	**0,033**
Invasão vascular venosa (sim)	0,060	-0,357	0,376	**0,139**	**0,008**
Recetor de estrogénio (positivo)	0,026	-0,241	0,267	**-0,161**	**0,001**
Recetor de progesterone (positivo)	0,033	-0,219	0,193	**0,081**	**0,015**
Triplo negativo (sim)	-0,002	-0,455	0,538	**-0,511**	**<0,0001**
Nódulos linfáticos (Nx ou N0 ou N1)	0,109	-0,496	0,817	**0,354**	**<0,0001**
Processo de Sobrevivência					
Idade ao diagnóstico (continua)	**-0,028**	**-0,057**	**-0,003**	**-0,046**	**<0,0001**
Estadio (III ou IV)	**1,584**	**0,844**	**2,243**	**1,399**	**<0,0001**
Triplo negativo (sim)	**2,110**	**0,728**	**3,362**	**1,946**	**<0,0001**
Associação Latente					
Parâmetros de Associação	$\gamma_0 =$ **0,922**	**0,175**	**1,694**	$\alpha =$ **1,677**	**<0,0001**
$\hat{\nu}^2$	**0,132**	-	-	**0,390**	-
$\hat{\tau}^2$	**0,012**	-	-	**0,113**	-
Loglikelihood	**-1744,953**			**-1726,397**	

A Tabela 4.18 apresenta o modelo conjunto considerado para o marcador tumoral CEA. São apresentadas as estimativas dos coeficientes, os intervalos de confiança associados no caso do *package joineR* e o p-valor no caso do *package JM*, os parâmetros de associação, a variância do processo aleatório e dos erros e a Loglikelihood dos modelos. Relativamente ao processo longitudinal, consegue-se identificar que há um maior número de variáveis estatisticamente significativas usando o *package JM* em comparação com o *joineR*. Isto deve-se à própria interpretação dos coeficientes no modelo. No processo de sobrevivência as variáveis apresentadas são estatisticamente significativas em ambos os modelos.

Examinando com mais detalhe o modelo conjunto obtido com o *package JM*, percebe-se que o processo longitudinal e o processo de sobrevivência estão associados visto que o valor do parâmetro de associação é estatisticamente significativo. Há evidência estatística para afirmar que existe uma variação ao longo do tempo para pacientes sem recidiva do cancro da mama, aumentando cerca de 0,037 unidades de log(CEA) por cada ano, e que pacientes com recidiva apesentam uma progressão do marcador mais acelerada aumentando 0,213 por cada ano, não sendo significativas as diferenças dos valores iniciais para pacientes com e sem recidiva da doença. Todas as restantes covariáveis, à exceção das variáveis menopausa, recetores de estrogénio e subtipo triplo negativo, levam a um aumento dos valores iniciais da progressão do marcador tumoral. No processo de sobre-

vivência, considerando-se a *hazard ratio* ($\exp(\beta)$), concluí-se que a idade ao diagnóstico tem um efeito protetor e as restantes variáveis representam um maior risco de se desenvolver a recidiva da doença. Este valor obtido tem em consideração a influência da parte longitudinal no processo de sobrevivência apresentado.

Analisando o modelo obtido com o *package joineR*, verifica-se que o parâmetro de associação é estatisticamente significativo, mostrando que a especificidade de cada sujeito parece influenciar a sua sobrevivência. Neste modelo a progressão do marcador varia ao longo do tempo, aumentando 0,038 por cada ano. Pacientes com recidiva, apesar de não apresentarem valores iniciais diferentes apresentam uma progressão dos valores do marcador mais acelerada, cerca de 0,179 por cada ano. Por cada ano a mais na idade de diagnóstico os valores do marcador tumoral CEA na escala logarítmica aumenta 0,017. No processo de sobrevivência, considerando-se a *hazard ratio* percebe-se que a idade ao diagnóstico tem um efeito protetor e as restantes variáveis representam um maior risco de recidivar a doença.

4.4.1.1 Análise de diagnóstico

Como referido no Capítulo 3, a análise de diagnóstico consiste na verificação de três pressupostos. No processo longitudinal verifica-se a normalidade de Z_{ij} e que a variância de Z_{ij} é constante. No processo de sobrevivência avalia-se a adequabilidade do modelo com base nos resíduos de Cox-Snell. Em seguida apresenta-se a análise de diagnóstico do modelo obtido com o *package JM*.

Analisando a Figura 4.52 verifica-se que a normalidade de Z_{ij} não parece ser rejeitada, apesar de haver um desvio em relação à reta. A Figura 4.53 como não apesenta nenhum padrão especifico mostra que o pressuposto da variância de Z_{ij} constante é verificado.

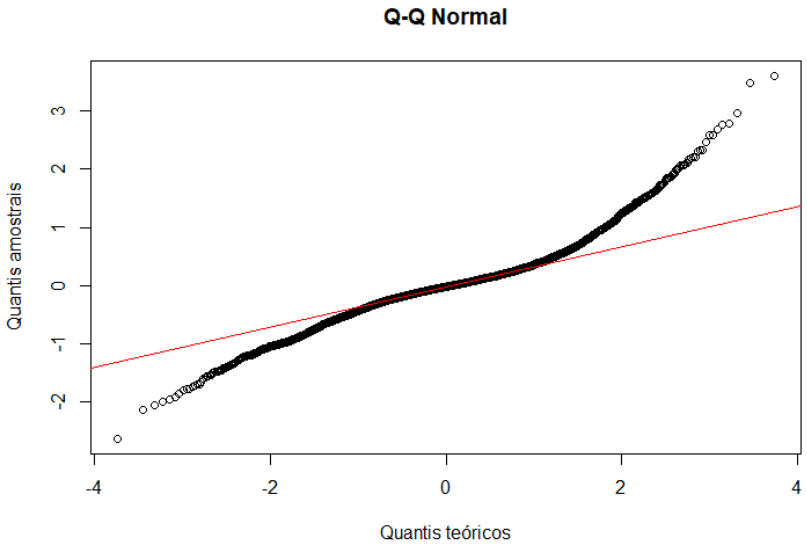

Figura 4.52: Q-Q normal dos resíduos: modelos conjuntos CEA

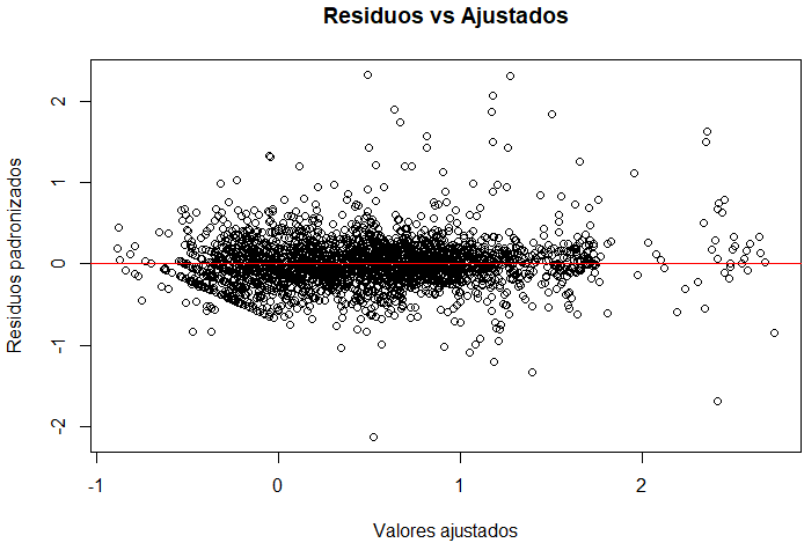

Figura 4.53: Resíduos *versus* valores ajustados: modelos conjuntos CEA

Como a representação dos resíduos de Cox-Snell *versus* a estimativa de Nelson-Aalen da função de risco cumulativa dos resíduos resultantes do modelo de sobrevivência (Figura 4.54) ajustado é aproximadamente uma reta de declive um e ordenada zero, concluí-se que o modelo é adequado visto que a amostra dos resíduos se pode considerar como proveniente de uma população exponencial de valor médio um.

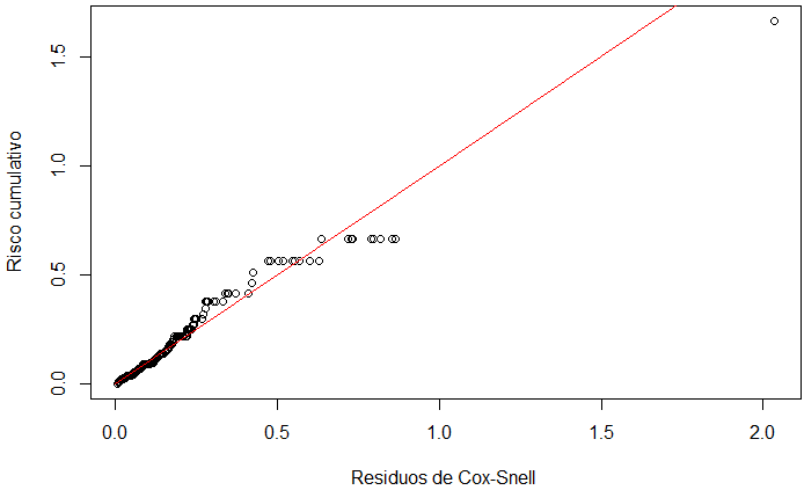

Figura 4.54: Resíduos Cox-Snell: modelos conjuntos CEA

4.4.2 Marcador tumoral CA15-3

O modelo conjunto do processo longitudinal dos valores do marcador tumoral CA15-3 e o processo de sobrevivência, foi obtido de forma semelhante à descrita na secção anterior

para o marcador tumoral CEA. Para a estimação dos modelos considerou-se 394 pacientes e um total de 3128 observações do marcador tumoral CA15-3.

Tabela 4.19: Modelos conjuntos *packages* joineR e JM: CA15-3

Modelos conjuntos	**joineR**			**JM**	
(sub-base de dados TR)	**Estimativa**	**Inferior 95%**	**Superior 95%**	**Estimativa**	**p-valor**
Processo Longitudinal					
Intercept	**1,699**	**0,969**	**2,171**	**1,830**	**<0,0001**
Tempo desde diagnóstico (anos)	0,011	-0,008	0,028	0,003	0,592
Tempo desde diagnóstico:Recidiva	0,044	-0,030	0,151	**0,098**	**<0,0001**
Idade ao diagnóstico (continua)	**0,007**	**0,003**	**0,013**	**0,003**	**0,016**
Bilateral (sim)	0,202	-0,117	0,495	**0,210**	**<0,0001**
Estadio (III ou IV)	0,356	-0,108	0,712	**0,265**	**<0,0001**
Recidiva (sim)	**0,348**	**0,128**	**0,716**	**0,511**	**<0,0001**
Mama (esquerda)	**-0,106**	**-0,272**	**-0,003**	**-0,198**	**<0,0001**
Presença de carcinoma associado (sim)	0,128	-0,003	0,250	**0,148**	**<0,0001**
Invasão vascular linfática (sim)	0,013	-0,176	0,187	**0,111**	**0,005**
Invasão vascular venosa (sim)	**0,538**	**0,098**	**1,064**	**0,632**	**<0,0001**
Recetor de estrogénio (positivo)	0,120	-0,147	0,339	**0,200**	**<0,0001**
Nódulos linfáticos (Nx ou N0 ou N1)	0,354	-0,169	0,827	**0,460**	**<0,0001**
Processo de Sobrevivência					
Idade ao diagnóstico (continua)	**-0,025**	**-0,055**	**-0,008**	**-0,030**	**<0,0001**
Grau (G3)	0,620	-0,228	1,384	**0,770**	**0,014**
Estadio (III ou IV)	**1,445**	**0,709**	**2,009**	**0,965**	**0,002**
Triplo negativo (sim)	**1,752**	**0,240**	**3,015**	**1,954**	**<0,0001**
Associação Latente					
Parâmetros de Associação	$\gamma_0 = 0{,}4441$	-0,759	0,936	$\boldsymbol{\alpha = 1{,}380}$	**<0,0001**
$\hat{\nu}^2$	**0,155**	-	-	**0,410**	-
$\hat{\tau}^2$	**0,056**	-	-	**0,236**	-
Loglikelihood	**-2986,032**			**-2959,90**	

A Tabela 4.19 apresenta o modelo conjunto considerado para o marcador tumoral CA15-3. São apresentadas as estimativas dos coeficientes, os intervalos de confiança associados no caso do *package joineR* e o p-valor no caso do *package JM*, os parâmetros de associação, a variância do processo aleatório e dos erros e a Loglikelihood dos modelos. Relativamente ao processo longitudinal conseguiu-se identificar que há um maior número de variáveis estatisticamente significativas usando o *package JM* em comparação com o *joineR*. Isto deve-se, como já foi referido, à própria interpretação dos coeficientes no modelo. No processo de sobrevivência as diferenças são menores havendo apenas uma variável que não é estatisticamente significativa.

Analisando com mais detalhe para o modelo conjunto obtido com o *package JM*, percebe-se que o processo longitudinal e o processo de sobrevivência estão associados visto que o valor do parâmetro de associação é estatisticamente significativo. Percebe-se que não parece existir uma variação ao longo do tempo para pacientes sem recidiva do cancro da mama, mas pacientes com recidiva apresentam valores iniciais mais elevados, cerca de 0,511, e uma progressão do marcador mais acelerado aumentando 0,098 por cada ano. Todas as restantes covariáveis, à exceção da variável que representa o lado da mama afetado, levam a um aumento dos valores iniciais da progressão do marcador tumoral. No processo de sobrevivência, considerando-se a *hazard ratio* ($\exp(\beta)$) percebe-se que a idade ao diagnóstico tem um efeito protetor e as restantes variáveis um maior risco de se desenvolver a recidiva da doença. Este valor obtido tem em consideração a influência da

parte longitudinal no processo de sobrevivência apresentado.

Analisando o modelo obtido com o *package joineR*, concluí-se que o parâmetro de associação estimado não é estatisticamente significativo. Este resultado é bastante interessante uma vez que permite concluir que a especificidade de cada sujeito não parece influenciar a sua sobrevivência. Neste modelo a progressão do marcador não parece variar ao longo do tempo, mas pacientes com recidiva tem valores iniciais mais elevados. Pacientes com cancro na mama esquerda parecem apresentar valores do marcador inferiores, e a presença de invasão vascular venosa traduz-se em valores tumorais mais elevados. No processo de sobrevivência, a idade ao diagnóstico mantêm um efeito protetor e pacientes com Estadio mais avançado e com subtipo triplo negativo apresentam maior risco de recidivarem a doença.

4.4.2.1 Análise de diagnóstico

À semalhança do que se fez com o marcador tumoral CEA, a análise de diagnóstico verifica os três pressupostos considerando o modelo obtido no *package JM*. No processo longitudinal verifica-se a normalidade de Z_{ij} e que a variância de Z_{ij} é constante. No processo de sobrevivência avalia-se a adequabilidade do modelo com base nos resíduos de Cox-Snell.

Analisando a Figura 4.55 verifica-se que a normalidade de Z_{ij} não parece ser rejeitada. A Figura 4.56 como não apesenta nenhum padrão específico mostra que o pressuposto da variância de Z_{ij} constante é verificado.

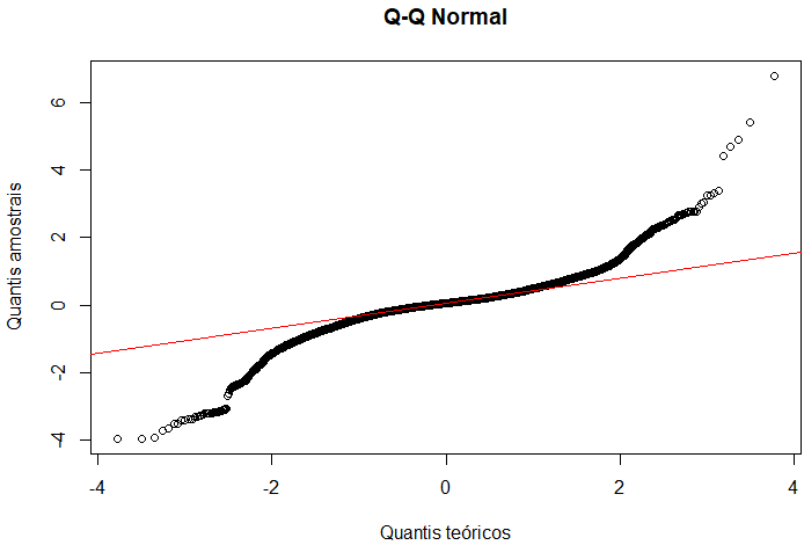

Figura 4.55: Q-Q dos resíduos: modelos conjuntos CA15-3

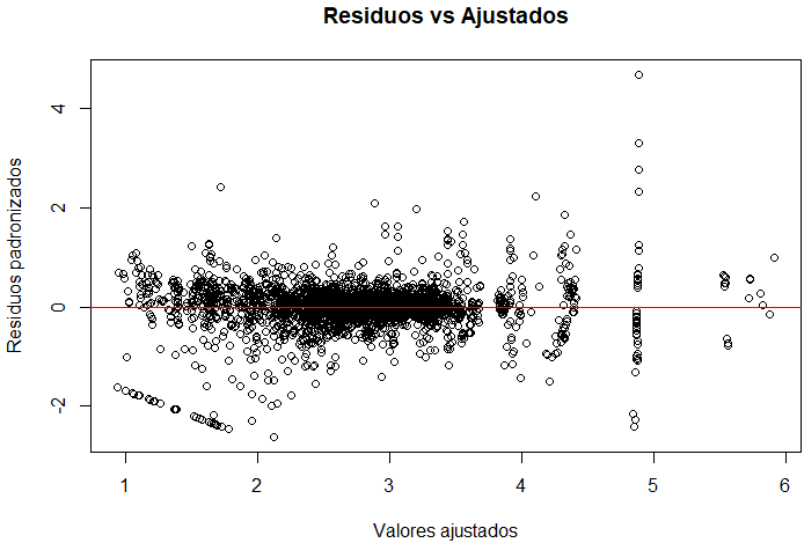

Figura 4.56: Resíduos *versus* valores ajustados: modelos conjuntos CA15-3

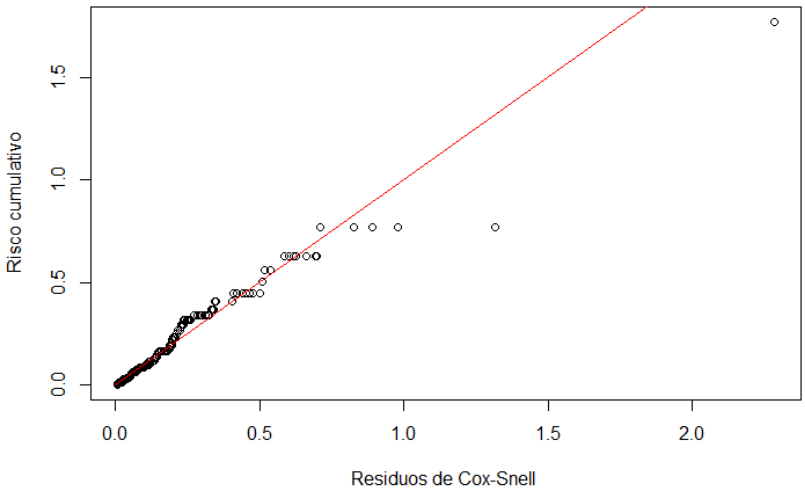

Figura 4.57: Resíduos Cox-Snell: modelos conjuntos CA15-3

Como na representação dos resíduos de Cox-Snell *versus* a estimativa de Nelson-Aalen da função de risco cumulativa dos resíduos resultantes do modelo de sobrevivência ajustado (Figura 4.57) a maioria dos pontos sobrepõem a recta de declive um e ordenada zero, concluí-se que o modelo é adequado visto que se pode considerar que a amostra dos resíduos é proveniente de uma população exponencial de valor médio um.

Capítulo 5

Conclusões

As diferentes análises efetuadas permitiram concluir que quando se utilizam bases de dados diferentes a significância das variáveis nos modelos é afetada. Quando, por exemplo, se compararam os resultados dos modelos na análise longitudinal tendo em conta ou não o total das observações conseguiu-se entender esta mesma diferença. A própria aborgadem de modelos conjuntos leva a conclusões distintas das análises realizadas em separado. Dada a significância dos parâmetros de associação das análises de modelos conjuntos, concluí-se que esta abordagem é mais adequada. Quer em termos práticos pela própria significância dos resultados, ao revelarem que a sobrevivência é influênciada pela progressão da doença, como em termos estatísticos no sentido em que os modelos parecem apresentar um melhor ajuste aos dados. Por exemplo, na análise de diagnóstico da análise longitudinal do marcador tumoral CEA, verifica-se que valores ajustados mais elevados estão associados a valores mais elevados dos resíduos. Na análise conjunta do mesmo marcador tumoral, esta relação parece desaparecer.

Relativamente aos resultados das análises, este estudo comprovou a necessidade do seguimento das pacientes com cancro da mama para a identificação adequada e atempada de uma recidiva. Há de facto alguns riscos acrescidos para a recidiva da doença em pacientes com determinadas características.

A análise de sobrevivência efetuada permitiu concluir que há um maior risco de recidiva do cancro da mama em pacientes com subtipo triplo negativo como referido por Rietjens et al.(2012). Pacientes com este subtipo de cancro têm cerca de 4,6 vezes maior risco de recidivar a doença. Estadios mais avançadas no diagnóstico apresentam cerca de 4,3 vezes maior risco de recidiva do cancro da mama e uma classificação Grau III no grau de Bloom e Richardson apresenta 2 vezes maior risco de recidiva da doença. A análise da idade ao diagnóstico permitiu identificar que o aumento da idade revela ser um efeito protetor e que de facto há um maior risco de recidiva em pacientes mais jovens. Esta informação deve ser tida em conta quer no acompanhamento como na identificação dos grupos de risco para a recidiva da doença como sugerido por Cianfrocca e Goldstein (2004).

A análise do marcador tumoral CEA permitiu identificar que a progressão deste mar-

cador varia ao longo do tempo para todas as pacientes, mas pacientes com recidiva para além de apresentarem valores iniciais deste marcador mais elevados tem uma progressão estatisticamente mais acelerada. Isto sugere que uma correta interpretação dos valores deste marcador desde o diagnóstico do cancro da mama pode identificar pacientes com maior probabilidade de recidiva da doença. Há um aumento dos valores do marcador com o aumento da idade e identifica-se que pacientes com subtipo triplo negativo apresentam valores iniciais mais baixos que a média e quando a paciente tem um tumor primário T3 ou T4 ou Tx apresenta valores iniciais mais elevados. Na análise conjunta efetuada concluí-se que o grau de Bloom e Richardson deixa de ser significativo na sobrevivência do paciente não revelando haver maior risco de recidiva. As restantes variáveis identificadas na análise de sobrevivência em separado mantêm a significância. Em ambos os modelos apresentados a progressão de pacientes com recidiva continua a ser diferente, mas os valores iniciais não são diferentes das pacientes sem recidiva. No modelo obtido pelo *package joineR* o subtipo triplo negativo perde a sua significância nos valores iniciais da progressão do marcador tumoral e em ambos os modelos o tumor primário perde a significância. No entanto, as variáveis estadio, grau de Bloom e Richardson, menopausa, presença de carcinoma associado, invasão linfática e venosa, recetores de estrogénio e progesterona e nódulos linfáticos afetam a progressão do marcador e o risco de recidiva da doença. Concluí-se ainda que a modelação conjunta é uma mais valia visto que os processos estão associados e, no modelo obtido com o *package joineR*, verifica-se que a especificidade de cada indivíduo interfere na sua probabilidade de recidiva.

A análise do marcador tumoral CA15-3 permitiu concluir que não existe uma variação estatisticamente significativa na progressão deste marcador quando a paciente não tem recidiva, mas pacientes com recidiva do cancro da mama tem valores iniciais mais elevados e uma progressão que varia ao longo do tempo. Mais uma vez, concluí-se que uma correta interpretação destes valores permite identificar uma possível recidiva da doença nas pacientes. Há um aumento dos valores do marcador com o aumento da idade e pacientes com invasão vascular venosa tem valores iniciais mais elevados. A análise conjunta permitiu entender que o grau de Bloom e Richardson só é significativo na parte de sobrevivência quando se opta pelo modelo do *package JM*. O modelo obtido com o *package joineR* revela que a idade e a invasão vascular venosa mantêm o efeito sobre os valores do marcador. Contudo, a especificidade do indivíduo não parece influenciar a probabilidade de recidivar a doença. O modelo obtido com o *package JM* permite inferir que os processos estão associados. A progressão continua a não ter variação significativa quando não há uma recidiva associada, mas revela uma aceleração da progressão quando há. Pacientes com recidiva da doença tem valores iniciais mais elevados. As variáveis cancro bilateral, estadio, mama afetada, presença de carcinoma associado, invasão vascular linfática, recetores de estrogénio e nódulos linfáticos mostram-se agora significativas.

Concluí-se que os processos estudados estão relacionados da mesma forma que os

processos longitudinais e de sobrevivência revelaram uma forte associação quando o evento de interesse era a morte no trabalho de Borges (2015).

Para trabalho futuro seria muito interessante fazer-se uma análise de sobrevivência considerando a truncatura para se obterem estimativas mais corretas e conseguir-se estimar modelos conjuntos tendo-se em conta a mesma truncatura no processo de sobrevivência.

Capítulo 6

Anexos

Anexo A - Variáveis exploratórias ao nível individual: variáveis categóricas

Variável	Categoria	Número de pacientes	Percentegam
Distrito de residência	Braga	539	96,42%
	Porto	4	0,72%
	Viana do Castelo	16	2,86%
Habilitações literárias	1- $\leq 4^{o}$ ano	276	49,37%
	2- $\leq 9^{o}$ ano	86	15,38%
	3- $\leq 12^{o}$ ano	26	4,65%
	4- Curso superior	42	7,51%
	NA	129	23,08%
Número de partos	0	41	7,33%
	1	65	11,63%
	2	132	23,61%
	3	52	9,30%
	4	22	3,94%
	> 4	42	7,51%
	NA	205	36,67%
Amamentação	Sim	209	37,39%
	Não	14	2,50%
	NA	336	60,11%
Menopausa	Sim	343	61,36%
	Não	195	34,88%
	NA	21	3,76%
Tipo de menopausa	Cirurgica	27	4,83%
	Não cirurgica	265	47,41%
	NA	267	47,76%
Grau de parentesco	1- Pai, mãe, filho (primeiro grau)	18	3,22%
	2- Irmãos e avós (segundo grau)	24	4,29%
	3- Tios, sobrinhos e bisavós (terceiro grau)	16	2,86%
	4- Primos e trisavós (quarto grau)	9	1,61%
	NA	492	88,01%
Substituição hormonal terapêutica	Sim	26	4,65%
	Não	69	12,34%
	NA	464	83,01%
Contraceção hormonal oral	Sim	114	20,39%
	Não	70	12,52%
	NA	375	67,08%

Anexo B - Variáveis exploratórias ao nível do tumor: variáveis categóricas I

Variável		Categorias	Número de pacientes	Percentagem
Tipo de Recidiva		0- Sem recidiva	476	85,15%
		1- Local	11	1,97%
		2- Metastizada	64	11,45%
		3- Local e metastizada	6	1,07%
		NA	2	0.36%
Tipo de cirurgia		0- Sem cirurgia	14	2,5%
		1- Cirurgia conservadora da mama	224	40,07%
		2- Mastectomia com ou sem dissecção dos linfonodos axilares	321	57,42%
Quimioterapia		Sim	369	66,01%
		Não	181	32,38%
		NA	9	1,61%
Radioterapia		Sim	370	66,19%
		Não	175	31,31%
		NA	14	2,5%
Terapia hormonal		Sim	478	85,51%
		Não	63	11,27%
		NA	18	3,22%
Estado do tumor (TNM)	Tumor primário (T)	TX	12	2,15%
		T1	268	47,94%
		T2	205	36,67%
		T3	29	5,19%
		T4	7	1,25%
		Tis	28	5,01%
		NA	10	1,79%
	Grau de disseminação para nódulos linfáticos regionais (N)	NX	30	5,37%
		N0	269	48,12%
		N1	162	28,98%
		N2	45	8,05%
		N3	33	5,90%
		NA	20	3,58%
	Presença de metástases distantes (M))	M0 - Sem metastases distantes	544	97,32%
		M1 - Com metastases distantes	4	0,72%
		NA	11	1,97%
Estadio		Estadio 0	27	4,83%
		Estadio I	193	34,53%
		Estadio II	213	38,10%
		Estadio III	98	17,53%
		Estadio IV	4	0,72%
		NA	24	4,29%

Anexo C - Variáveis exploratórias ao nível do tumor: variáveis categóricas II

Variável	Categoria	Número de pacientes	Percentegam
Tipo Histórico do cancro	Ductal in situ	40	7,16%
	Ductal invasor	359	64,22%
	Invasor misto lobular	24	4,29%
	In situ lobular	6	1,07%
	Invasor misto lobular	41	7,33%
	Outro invasor	55	9,84%
	Outros	4	0,72%
	NA	30	5,37%
Invasão linfática	Sim	98	17,53%
	Não	461	82,47%
Invasão venosa	Sim	26	4,65%
	Não	533	95,35%
Grau de diferenciação- Bloom Richardson	Gx - não foi possível classificar o grau	2	0,36%
	G1 - carcinoma bem diferenciado	168	30,05%
	G2 - carcinoma moderadamente diferenciado	224	40,07%
	G3 - carcinoma pouco diferenciado	118	21,11%
	NA	47	8,41%
Recetores de estrogénio	Positivo	416	74,42%
	Negativo	66	11,81%
	NA	77	13,77%
Recetores de progesterona	Positivo	345	61,72%
	Negativo	114	20,39%
	NA	100	17,89%
Her2-neu	Positivo	247	44,19%
	Negativo	168	30,05%
	NA	144	25,76%
Triplo Negativo	Sim	21	3,76%
	Não	459	82,11%
	NA	79	14,13%
Ki-67	Alto	178	31,84%
	Baixo	136	24,33%
	NA	245	43,83%

Anexo D - Variáveis exploratórias ao nível individual: variáveis contínuas

Variável	Número de pacientes	Mínimo	1º Quartil	Mediana	Média	3º Quartil	Máximo	Desvio-padrão	NA(%)
Idade ao diagnóstico	599	20	47	58	58,69	71	92	14,25	-
Menarca	321	9	12	13	13,51	15	19	1,87	238(42,58)
Idade na primeira gravidez	224	15	22	25	25,33	28	40	4,65	335(59,92)
Idade da menopausa	188	33	45	50	48,63	52	58	58	371(66,37)

BIBLIOGRAFIA

American cancer society (2019). About breast Cancer. Consultado a 7 de Maio de 2021, retirado de *https://www.cancer.org/cancer/breast-cancer/about/what-is-breast-cancer.html*.

Benjamini, Y., e Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Series B Stat Methodol*.

Bloom, H. e Richardson, W. (1957). Histological grading and prognosis in breast cancer; a study of 1409 cases of which 359 have been followed for 15 years. *British Journal of Cancer*, **11**, 359-77.

Borges, A. (2015). Joint Modelling of Longitudinal and Survival Data on Breast Cancer. (Tese de Doutoramento, Universidade do Minho, Braga, Portugal). Obtido de *http://repositorium.sdum.uminho.pt/handle/1822/40426*

Cianfrocca, M. & Goldstein, L. (2004). Prognostic and predictive factors in earlystage breast cancer. *The Oncologist*, **6**.

Collett, D. (2003). *Modelling Survival Data in Medical Research*. 2nd edition, Chapman & Hall/CRC, Boca Raton.

Cox, D. R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society B*, **34**, 187-220.

Cox, D. R. e Snell, E. J. (1968). A general definition of residuals (with discussion). *Journal of the Royal Statistical Society, Series A*, **30**, 248-275.

de Oliveira, C.F. & da Silva, T.S. (2011). *Manual de Ginecologia*, vol. II, chap. 37. Permanyer Portugal.

Diggle, P. J. (1988). An approach to the analysis of repeated measurements, *Biometrics*, **44**, 959–971.

Diggle, P. J. (1998). *Dealing with missing values in longitudinal studies*. In "Recent Advances in the Statistical Analysis of Medical Data" (Everitt, B. S. and Dunn, G., Eds.), Arnold, London, 203–228.

Diggle, P. J., Heagerty, P., Liang, K-Y e Zeger, S. L. (2002). *Analysis of Longitudinal Data* . 2nd edition, Oxford: Oxford University Press.

DGS (2012). Recomendações Nacionais para Diagnóstico e Tratamento do Cancro da Mama. Consultado a 9 de Junho de 2021, retirado de *https://www.dgs.pt/documentos-e-publicacoes/ recomendacoes-nacionais-para-diagnostico-e-tratamento-do-cancro-da-mama.aspx*

DGS (2014). Portugal - Doenças Oncológicas em Números 2015. Consultado a 9 de Junho de 2021, retirado de *https://www.dgs.pt/estatisticas-de-saude/estatisticas-de-saude/publicacoes/ portugal-doencas-oncologicas-em-numeros-2014.aspx*

DGS (2015). Portugal - Doenças Oncológicas em Números 2015. Consultado a 9 de Junho de 2021, retirado de *https://www.dgs.pt/em-destaque/portugal-doencas -oncologicas-em-numeros-201511.aspx*

DGS (2017). Programa Nacional Para as Doenças Oncológicas 2017. Consultado a 9 de Junho de 2021, retirado de *https://www.iccp-portal.org/system/files/plans/DGS_ PNDO2017_ V10.pdf*

Fitzgibbons, P., Page, D., Weaver, D., Thor, A., Allred, D. e Clark, G. (2000). Prognostic factors in breast cancer, college of american pathologists consensus statement 1999. *Archives of Pathology & Laboratory Medicine*, **124**, 966-978.

Gehan, E.A. (1965). A generalized Wilcoxon test for comparing arbitrarily singly censored samples. *Biometrika*, **52**, 203-223.

Grambsch P.M. e Therneau, T.M. (1994). Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*, **81**, 515-526.

Gregoire, T. D., Brillinger, D. B., Diggle P. J., Russek -Cohen, E., Warren, W. G. e Wolfinger, R. D. (1997). *Modelling Longitudinal and Spatially Correlated Data.* Springer, 392-395.

Guadagni, F., Ferroni, P., Carlini, S., Mariotti, S., Spila, A., Aloe, S., D'Alessandro, R., Carone, M., Cicchetti, A., Ricciotti, A., Venturo, I., Perri, P., Filippo, F.D., Cognetti, F., Botti, C. e Roselli, M. (2001). A re-evaluation of carcinoembryonic antigen (cea) as a serum marker for breast cancer: A prospective longitudinal study. *Clinical Cancer Research*, **7**, 2357-2362.

Henderson, R., Diggle, P. e Dobson, A. (2000). Joint modelling of longitudinal measure-

ments and event time data. *Biostatistics*, **1**, 465-480.

Hogan, J. W. e Laird, N. M. (1997). Model-based approaches to analysing incomplete longitudinal and failure time data, *Statistics in Medicine*, **16**, 259–272.

Hospital de Braga EPE (2020), O Hospital. Consultado a 18 de Maio de 2021, retirado de
https://www.hospitaldebraga.pt/hospital/sobre-nos.

Hwang, K-T, Kim, Y. A., Kim, J., Chu, A. J., Chang, J. H., Oh, S. W., Hwang, K. R., Chai, Y. J. (2017). The influences of peritumoral lymphatic invasion and vascular invasion on the survival and recurrence according to the molecular subtypes of breast cancer. *Springer Science+Business Media.* New York.

Kaplan, E. L. e Meier, P. (1958), Nonparametric estimation from incomplete observations, *Journal of the American Statistical Society* , **53**, 457-481.

Klein, J.P e Moeschberger, M.L. (1997). *Survival Analysis: Techniques for Censored and Truncated Data.* New York: Springer.

Liga Portuguesa contra o cancro (s.d), Cancro da mama. Consultado a 5 de Maio de 2021, retirado de *https://www.ligacontracancro.pt/cancro-da-mama/*

Little, R. J. e Rubin, D. B. (2000). *Statistical Analysis with Missing Data.* 2nd edition, New York: Wiley.

McCrink, L., Marshall, A. e Cairns,K. (2013). Advances in joint modelling: A review of recent developments with application to the survival of end stage renal disease patients. *International Statistical Review*, **81**, 249-269.

Philipson, P., Sousa, I., Diggle, P., Williamson, P., Kolamunnage-Dona, R., Henderson, R. e Team, R.C. (2012). Joiner: Joint modelling of repeated measurements and time-to-event data.

Peto, R. e Peto, J. (1972). Asymptotically efficient rank invariant test procedures (with discussion). *Journal of the Royal Statistical Society, Series A*, **135**, 185-206.

Rietjens M., Pedrini, J. L.,Lohsiriwat, V., Schorr, M. C., Zetler, C., Reginatto, A. G., Gonzalez, T., Bacco, R. (2012). Recidiva locorregional: importância da margem cirúr-

gica livre e dos subtipos moleculares do câncer de mama. *Revista Brasileira de Mastologia.*

Rizopoulos, D. (2010). Jm: An r package for the joint modelling of longitudinal and time-to-event data. *Journal of Statistical Software*, **35**, 1-33 .

Rocha, C. e Papoila, A. L. (2009). *Análise de Sobrevivência.* Sociedade Portuguesa de Estatística.

Rodrigues, V. (2011). *Manual de Ginecologia*, vol. **II**, chap. 34, 175-201. Permanyer Portugal.

Schoenfeld, D.A. (1982). Partial residuals for the proportional hazards regression model. *Biometrika*, **69**, 239-241

Sousa, I. (2011). A review on joint modelling of longitudinal measurements and time-to event. *REVSTAT*, **9**, 57-81.

Trichopoulos, D., Adami, H., Ebkom, A., Hsieh, C. & Lagiou, P. (2008). Early life events and conditions and breast cancer risk: from epidemioloy to etiology. *International Journal of Cancer*, **122**, 481-495.

Tsai, W., Jewell, N. P. & Wang, M. (1987). A note on the product-limit estimator under right censoring and left truncation, *Biometrika*, **74**, 883-886.

Vasconcelos, A. L., Ferreira, A. R., Sousa, B., Loewenthal, C. S., Pinto, D., Cardoso, F., Gervásio, H., Pereira, H., Vendrell, I., Coelho, J. L., Ribeiro, J. M., Marques, J. C, Paulo, J. V., Costa, L., Travado, L., Cardoso, M. J., Batista, M. V., Dionísio, M. R., Afonso, N., Gouveira, P., André, S., Castedo, S., Braga, S. A., Pedro, S. & Matias, T. (2017) *100 perguntas chave no cancro da mama.* 2º Edição, Permanyer Portugal.

Wang, M. (1991). Nonparametric estimation from cross-sectional survival data, *Journal of the American Statistical Association* vol. **86**, no. 413, 130-143.

Wang, M., Broojmeyer, R. e Jewell, N. P. (1993). Statistical models for prevalent cohort data, *Biometrics*, **49**, 1-11

Wulfsohn, M. e Tsiatis, A. (1997). A joint model for survival and longitudinal data measured with error. *Biometrics*, **53**, 330-339.