LEONARDO FREITAS BOAVENTURA RIOS

**Modelos de predição de risco de morte para pacientes
com carcinoma epidermoide de cabeça e pescoço**

Tese apresentada à Faculdade de Medicina da
Universidade de São Paulo para obtenção do
título de Doutor em Ciências

Programa de Ciências Médicas
Área de concentração: Distúrbios do
Crescimento Celular, Hemodinâmicos e da
Hemostasia
Orientador: Dra. Raquel Ajub Moyses

São Paulo

2020

LEONARDO FREITAS BOAVENTURA RIOS

**Modelos de predição de risco de morte para pacientes
com carcinoma epidermoide de cabeça e pescoço**

Tese apresentada à Faculdade de Medicina da
Universidade de São Paulo para obtenção do
título de Doutor em Ciências

Programa de Ciências Médicas
Área de concentração: Distúrbios do
Crescimento Celular, Hemodinâmicos e da
Hemostasia
Orientador: Dra. Raquel Ajub Moyses

São Paulo

2020

Dados Internacionais de Catalogação na Publicação (CIP)

Preparada pela Biblioteca da
Faculdade de Medicina da Universidade de São Paulo

©reprodução autorizada pelo autor

Rios, Leonardo Freitas Boaventura
Modelos de predição de risco de morte para
pacientes com carcinoma epidermoide de cabeça e
pescoço / Leonardo Freitas Boaventura Rios. -- São
Paulo, 2020.

Tese(doutorado)--Faculdade de Medicina da
Universidade de São Paulo.

Programa de Ciências Médicas. Área de
Concentração: Distúrbios do Crescimento Celular,
Hemodinâmicos e da Hemostasia.

Orientadora: Raquel Ajub Moyses.

Descritores: 1.Neoplasia de cabeça e pescoço
2.Mineração de dados 3.Predição 4.Modelos logísticos
5.Redes neurais 6.Árvores de decisões 7.Análise de
sobrevida 8.Prognóstico

USP/FM/DBD-016/20

Responsável: Erinalva da Conceição Batista, CRB-8 6755

Aos meus pais, **Adenor e Fátima**, duas fontes inesgotáveis de amor e empenho. Dedico a vocês por me ensinar a acreditar nos sonhos mais distantes através do estudo: “estudar, estudar e estudar”.

AGRADECIMENTOS

À Deus, nosso criador e motivo da nossa existência.

Aos nossos pacientes. Espero que estes dados de alguma forma os ajudem ao longo do tratamento.

À minha esposa Ruth, fonte de inspiração diária. Tenho muita sorte de poder contar contigo. Obrigado pelo companheirismo e amor ao longo desta jornada. Meus filhos Pedro e Liz, pelos abraços e sorrisos contagiantes toda vez que eu voltava de viagem.

À minha orientadora Dra. Raquel Ajub Moysés pela inestimável importância no meu crescimento científico, profissional e pessoal. Obrigado por ter acreditado em mim e por tudo que aprendi ao longo deste estudo. Agradeço também ao Ronaldo Frizzarini e ao Henrique pelos momentos em que a Raquel teve que se ausentar, principalmente nos últimos seis meses.

Ao Prof. Dr. Pedro Michaluart Jr, por ter aberto as portas do LIM-28 para mim. Jamais esquecerei do dia em que conversávamos a respeito de ferramentas preditivas e ele me mostrou o projeto temático do GENCAPO. Professor, seu legado continuará se multiplicando.

Às minhas irmãs queridas Carollina e Fernanda pelo carinho e amor existente entre nós.

À minha madrinha e avó Maria. Uma grande apoiadora, incentivadora e torcedora ao longo de toda minha jornada como estudante, sobretudo ao longo da graduação em medicina.

A equipe do Laboratório LIM-28, em especial minha amiga Pamela. Obrigado pelo auxílio nos momentos que eu estava na Bahia.

A todos os participantes do projeto GENCAPO. O empenho de vocês na coleta dos dados tornou este estudo possível.

Aos membros da banca de qualificação, Prof. Dr. Gilberto de Castro Jr, Prof. Dr. Leandro Luongo de Matos, Prof. Dra. Tatiana Natasha Toporcov e Dr. Marcos Brasilino pelas sugestões e opiniões.

Ao Dr. José Eduardo Levi, Dr. Rossana Verónica Mendoza López e Dr. Alexandre Bezerra pelas valiosas informações a respeito do HPV.

À Keila Taira Machado e Rafael Pereira de Souza por toda ajuda no levantamento dos dados faltantes do Hospital Heliópolis e Instituto do Câncer Arnaldo Vieira de Carvalho, respectivamente.

Ao Prof. Dr. Lenine Garcia Brandão, professor titular do Departamento de Cirurgia de Cabeça e Pescoço do HC-FMUSP no início deste projeto.

À Dra. Marília D'Elboux Guimarães Brescia, Dr. Caio Plopper, Dr. Caio Tosato Caliseo, Dr. Jorge Du Ub Kim, por toda a ajuda no levantamento dos questionários e dados faltantes dos pacientes do Hospital das Clínicas da Faculdade de Medicina da Universidade de São Paulo.

À toda equipe da Fundação Oncocentro de São Paulo (FOSP) pelo trabalho exemplar em cortar os blocos e realizar a pesquisa do HPV. Em especial a Diane Dede Cohen, Ricardo Fernandes Goes e Luiz Mario Santos Fuza.

À Universidade Estadual de Feira de Santana, por ter permitido minha ausência ao longo dos últimos três anos, em especial o Dr. Jose de Bessa Jr e Dr. Edval Gomes dos Santos Junior.

A todos os amigos da Santa Casa de Misericórdia de Feira de Santana pelo convívio ao longo das quartas-feiras e por me representar tão bem quando estou ausente. Em especial ao Dr. Bruno Pinto Ribeiro, Clériston Nunes, Tanielle Falcão e Rebeca Nunes.

Às secretárias da pos graduação Eliane (Clínica Cirúrgica) e Angélica (Ciências Médicas) pela ajuda ao longo de toda esta jornada.

Normas adotadas neste estudo:

Referências: adaptado de International Committee of Medical Journal Editors
(Vancouver).

Universidade de São Paulo. Faculdade de Medicina. Divisão de Biblioteca e Documentação. Guia de apresentação de dissertações, teses e monografias. Elaborado por Anneliese Carneiro da Cunha, Maria Júlia de A. L. Freddi, Maria Fazanelli Crestana, Marinalva de Souza Aragão, Suely Campos Cardoso, Valéria Vilhena. 3ª edição. São Paulo: Divisão de Biblioteca e Documentação; 2011.

Abreviaturas dos títulos dos periódicos de acordo com List of Journals Indexed in Index Medicus.

Novo Acordo Ortográfico da Língua Portuguesa (Decreto no 6.583 de 29/09/2008)

Descritores em Ciências da Saúde - DeCS (<http://decs.bvs.br/>)

"A vida é o que fazemos dela. As viagens são os viajantes. O que vemos não é o que vemos, senão o que somos." (Fernando Pessoa)

SUMÁRIO

Lista de siglas

Lista de abreviaturas e símbolos

Lista de figuras

Lista de tabelas

Resumo

Abstract

1.	INTRODUÇÃO	1
2.	OBJETIVO	4
3.	REVISÃO DA LITERATURA	6
3.1.	Epidemiologia – carcinoma epidermoide no Brasil e no mundo	7
3.2.	TNM	8
3.2.1.	Incorporação do HPV ao TNM	10
3.2.2.	Fatores prognósticos não contemplados pelo TNM em CECp	10
3.2.2.1.	Tabagismo	11
3.2.2.2.	Raça	11
3.2.2.3.	Idade	12
3.3.	Predição em outras áreas da oncologia	12
3.4.	Novos conceitos	13
3.5.	Desenvolvimento de modelos preditivos em oncologia	13
3.5.1.	Coleta de Dados	14
3.5.2.	Análise de consistência	15
3.5.3.	Exploração de dados	15
3.5.4.	Seleção de variáveis	15
3.5.4.1.	Categorização de variáveis	16
3.5.4.2.	Análise de componentes principais	16
3.5.5.	Aditividade dos efeitos preditivos e interações entre as variáveis	17
3.5.6.	Dados faltantes	18
3.5.6.1.	Aleatoriedade de dados faltantes	19
3.5.6.2.	Recomendações e limitações da imputação de dados	19
3.5.7.	Ajustes para análises múltiplas	20
3.5.8.	Métodos estatísticos para modelos preditivos	20
3.5.8.1.	Regressões	20
3.5.8.1.1.	Seleção de variáveis em modelos de regressão	22
3.5.8.2.	Redes Neurais	23
3.5.8.3.	Métodos baseados em árvores	24
3.5.8.3.1.	Árvores de regressão e classificação	24
3.5.8.3.2.	Floresta Aleatória	27
3.5.8.3.3.	<i>Boosted regression trees</i>	28
3.5.8.4.	Modelos Dinâmicos	28
3.5.9.	Validação de métodos	29
3.5.10.	Ajuste	32
3.5.11.	Testando a performance de um modelo preditivo	33
4.	MÉTODOS	35
4.1.	Seleção dos indivíduos	36
4.2.	Ética	36
4.3.	Preenchimento dos questionários Gencapo	37

4.4.	Seleção das variáveis	37
4.5.	Material biológico	39
4.6.	Pesquisa do Gene do Papilomavírus Humano	39
4.7.	Extração de DNA de tecido parafinado	40
4.8.	Métodos estatísticos	41
4.8.1.	Imputação de Dados	41
4.8.1.1.	Dimensões Tumorais no Estudo de Anatomia Patológica	41
4.8.1.2.	Dimensões Tumorais Clínicas	42
4.8.1.3.	Sítios dos tumores primários	42
4.8.1.4.	Perda Ponderal (meses)	44
4.8.2.	Análise Estatística	45
4.8.3.	Validação	46
4.8.4.	Transformação de variáveis contínuas	46
4.8.4.1.	Idade	46
4.8.4.2.	Perda ponderal	47
4.8.4.3.	Volume clínico	47
4.8.5.	Criação da variável linfonodomegalia	49
4.9.	Variáveis utilizadas	49
5.	RESULTADOS	50
5.1.	Descrição da amostra total de pacientes	52
5.2.	Modelos preditivos para tumores de todos os sítios	55
5.2.1.	Modelos baseados em árvores	55
5.2.1.1.	Modelo CART para tumores de todos os sítios	55
5.2.1.2.	<i>Boosted Regression Trees</i> para tumores de todos os sítios	57
5.2.1.3.	Floresta Aleatória para tumores de todos os sítios	59
5.2.2.	Redes Neurais para tumores de todos os sítios	60
5.2.3.	Regressão Logística para tumores de todos os sítios	61
5.2.4.	Calibração dos modelos construídos para tumores de todos os sítios	64
5.2.5.	Comparação dos modelos construídos para tumores de todos os sítios com o TNM	65
5.3.	Modelos preditivos para tumores de cavidade oral	65
5.3.1.	Modelos baseados em árvores	65
5.3.1.1.	Modelo CART para tumores de cavidade oral	65
5.3.1.2.	<i>Boosted Regression Trees</i> para tumores de cavidade oral	67
5.3.1.3.	Floresta Aleatória para tumores de cavidade oral	68
5.3.2.	Redes Neurais para tumores de cavidade oral	70
5.3.3.	Regressão Logística para tumores de cavidade oral	71
5.3.4.	Calibração dos modelos construídos para tumores de cavidade oral	73
5.3.5.	Comparação dos modelos construídos para tumores de cavidade oral com o TNM	74
5.4.	Modelos preditivos para tumores de orofaringe	74
5.4.1.	Modelos baseados em árvores	74
5.4.1.1.	Modelo CART para tumores de orofaringe	74
5.4.1.2.	<i>Boosted Regression Trees</i> para tumores de orofaringe	76
5.4.1.3.	Floresta Aleatória para tumores de orofaringe	78
5.4.2.	Redes Neurais para tumores de orofaringe	79
5.4.3.	Regressão Logística para tumores de orofaringe	80
5.4.4.	Calibração dos modelos construídos para tumores de orofaringe	82

5.4.5.	Comparação dos modelos construídos para tumores de orofaringe com o TNM	83
5.5.	Modelos preditivos para tumores de hipofaringe	84
5.5.1.	Modelos baseados em árvores	84
5.5.1.1.	Modelo CART para tumores de hipofaringe	84
5.5.1.2.	<i>Boosted Regression Trees</i> para tumores de hipofaringe	86
5.5.1.3.	Floresta Aleatória para tumores de hipofaringe	87
5.5.2.	Redes Neurais para tumores de hipofaringe	89
5.5.3.	Regressão Logística para tumores de hipofaringe	90
5.5.4.	Calibração dos modelos construídos para tumores de hipofaringe ...	91
5.5.5.	Comparação dos modelos construídos para tumores de hipofaringe com o TNM	92
5.6.	Modelos preditivos para tumores de laringe	93
5.6.1.	Modelos baseados em árvores	93
5.6.1.1.	Modelo CART para tumores de laringe	93
5.6.1.2.	<i>Boosted Regression Trees</i> para tumores de laringe	95
5.6.1.3.	Floresta Aleatória para tumores de laringe	96
5.6.2.	Redes Neurais para tumores de laringe	98
5.6.3.	Regressão Logística para tumores de laringe	99
5.6.4.	Calibração dos modelos construídos para tumores de laringe	101
5.6.5.	Comparação dos modelos construídos para tumores de laringe com o TNM	101
5.7.	Incremento da performance dos modelos preditivos após adição do status HPV	102
5.7.1.	Tumores de todos os sítios	102
5.7.2.	Tumores de orofaringe	103
6.	DISCUSSÃO	105
6.1.	Limitações	115
6.2.	Perspectivas futuras	117
7.	CONCLUSÃO	119
8.	ANEXOS	121
9.	REFERÊNCIAS	130
Apêndice		

LISTA DE SIGLAS

AIC	<i>Akaike information criterion</i> (Critério de informação Akaike)
AP	Anatomia Patológica
ASC	Área sob a curva
BIC	<i>Bayesian information criterion</i> (Critério de informação Bayesiano)
BRCA	<i>Breast cancer</i> (câncer de mama)
BRT	<i>Boosted regression trees</i>
CAPPesq	Comissão de Ética para Análise de Projetos de Pesquisa
CART	<i>Classification and regression trees</i> (Árvores de classificação e regressão)
CECP	Carcinoma epidermoide de cabeça e pescoço
CID	Código internacional de doenças
cN	Estadiamento clínico dos linfonodos regionais
Ct	<i>Cycle Threshold</i> (Limiar do ciclo)
cT	Estadiamento clínico do tumor primário
DC1	Maior dimensão Clínica (comprimento)
DC2	Menor dimensão clínica (largura)
DOI	<i>Depth of invasion</i> (profundidade de invasão)
DNA	Ácido desoxirribonucleico
F	Feminino
FAPESP	Fundação de Amparo à Pesquisa do Estado de São Paulo
FMUSP	Faculdade de Medicina da Universidade São Paulo
GLOBOCAN	<i>Cancer incidence and mortality worldwide</i>
HC-FMUSP	Hospital das Clínicas da Faculdade de Medicina da Universidade São Paulo

HPV	<i>Human Papillomavirus</i> (Papilomavírus humano)
IARC	<i>International Agency for Research on Cancer</i>
IC	Intervalo de confiança
ICESP	Instituto do Câncer do Estado de São Paulo
ICt	<i>Internal Control</i> (Controle Interno)
IMC	Índice de Massa Corpórea
INCA	Instituto Nacional do Câncer
LA	Laringe
LIM	Laboratório de Investigação Médica
M	Masculino
OF	Orofaringe
OF&CO	Orofaringe e cavidade oral
OR	<i>Odds ratio</i> (Razão de chances)
PCA	<i>Principal component analysis</i> (Análise de componentes principais)
PCR	<i>Polymerase chain reaction</i> (reação em cadeia de polimerase)
RC	Razão de Chances
RR	Risco relativo
ROC	<i>Receiver operating characteristic</i>
S/N	Sim ou não
TNM	Sistema TNM de classificação de tumores malignos
UICC	Union for International Cancer Control
USP	Universidade de São Paulo

LISTA DE ABREVIATURAS E SÍMBOLOS

cm centímetro(s)

cm³ centímetro cúbico

et al. e outros

mm milímetro(s)

μm micrômetro(s)

vs *versus* – neste caso, significando: em comparação a

% percentagem

= igual

+

mais

- menos

> maior que

≥ maior ou igual que

< menor que

≤ menor ou igual que

LISTA DE FIGURAS

Figura 1 - Parâmetros utilizados para definir o estágio do sítio primário no TNM (cavidade oral, faringe e laringe)	10
Figura 2 - Equação modelo na regressão logística	22
Figura 3 - Diagrama representando o modelo CART	25
Figura 4 - Crescimento da árvore	26
Figura 5 - Modelo CART para predição de readmissão hospitalar	27
Figura 6 - Exemplo de esquema de validação cruzada em 10 subgrupos (k=10) .	31
Figura 7 - Representação gráfica do método de reamostragem <i>Bootstrapping</i> com o objetivo de obter uma distribuição normal de subamostras	32
Figura 8 - Representação gráfica de um modelo sobreajustado	33
Figura 9 - Distribuição pelos sítios antes da realocação do Cavidade Oral/Orofaringe.....	42
Figura 10 - CART analisando o sítio e o desfecho, a fim de melhor alocar os casos sem definição de sítio entre cavidade oral e orofaringe	43
Figura 11 - Distribuição pelos sítios após a realocação	43
Figura 12 - Regressão linear bivariada	44
Figura 13 - Distribuição da Perda Ponderal (m)	45
Figura 14 - CART utilizado para a criação de categorias para a variável “Idade”	47
Figura 15 - Distribuição para idade categórica	47
Figura 16 - Comportamento do R^2 ao longo da sequência de divisões nas diferentes subdivisões da amostra de pacientes	56
Figura 17 - Árvore de classificação com todos os sítios, modelo CART	56
Figura 18 - Árvore de classificação e regressão – sítio primário cavidade oral	66
Figura 19 - Árvore de classificação e regressão – sítio primário orofaringe	75
Figura 20 - Árvore de classificação e regressão – sítio primário hipofaringe	85
Figura 21 - Árvore de classificação e regressão – sítio primário laringe	94

LISTA DE TABELAS

Tabela 1 - Estimativa do número de casos novos segundo sexo, por região.....	8
Tabela 2 - Definição de termos comumente usados na predição moderna	14
Tabela 3 - Número e proporção dos dados faltantes por variáveis	39
Tabela 4 - Categorização da variável “volume clínico” (valores em cm ³)	48
Tabela 5 - Comparação da distribuição entre “Volume Clínico” e “Volume Clínico LOG”	48
Tabela 6 - Distribuição de acordo com o sítio, cT, cN, M e o grau de diferenciação tumoral dos pacientes incluídos e excluídos das análises	53
Tabela 7 - Distribuição dos tamanhos tumorais em três dimensões	54
Tabela 8 - Distribuição do volume tumoral (em cm ³)	54
Tabela 9 - Medidas de R ² e número de pacientes para cada subdivisão da amostra - modelo CART para todos os sítios tumorais	55
Tabela 10 - Curvas ROC do modelo CART para todos os sítios	57
Tabela 11 - Contribuição das variáveis no modelo CART para todos os sítios	57
Tabela 12 - Contribuição das variáveis preditivas para construção do modelo de BRT para todos os sítios	58
Tabela 13 - Curvas ROC do modelo de BRT para todos os sítios	58
Tabela 14 - Contribuição das variáveis preditivas para construção do modelo de Floresta Aleatória para todos os sítios	59
Tabela 15 - Curvas ROC do modelo de Floresta Aleatória para todos os sítios	60
Tabela 16 - Curvas ROC do modelo de redes neurais para todos os sítios	60
Tabela 17 - Coeficientes estimados pela regressão logística: todos os sítios	62
Tabela 18 - Curvas ROC do modelo de regressão logística para todos os sítios ...	62
Tabela 19 - Razão de chances de cada variável para o modelo com todos os sítios	63
Tabela 20 - Teste " <i>goodness-of-fit</i> " – modelo de regressão logística para todos os sítios	64
Tabela 21 - Teste " <i>goodness-of-fit</i> " – modelo de redes neurais para todos os sítios	64
Tabela 22 - Curvas de sobrevida em cinco anos para tumores de todos os sítios ..	65

Tabela 23 - Medidas de R^2 e número de pacientes para subdivisão da amostra – modelo CART para tumores de cavidade oral	66
Tabela 24 - Curvas ROC do modelo de CART para cavidade oral	67
Tabela 25 - Contribuição das variáveis no modelo de CART para tumores de cavidade oral	67
Tabela 26 - Contribuição das variáveis preditivas para construção do modelo de BRT para tumores de cavidade oral	68
Tabela 27 - Curvas ROC do modelo de BRT para tumores de cavidade oral	68
Tabela 28 - Contribuição das variáveis preditivas para construção do modelo de Floresta Aleatória para tumores de cavidade oral	69
Tabela 29 - Curvas ROC do modelo de Floresta Aleatória para tumores de cavidade oral	70
Tabela 30 - Curvas ROC do modelo de redes neurais para tumores de cavidade oral	70
Tabela 31 - Coeficientes estimados pela regressão logística: cavidade oral	71
Tabela 32 - Razão de chances de cada variável para o modelo com tumores de cavidade oral	72
Tabela 33 - Curvas ROC do modelo de regressão logística para tumores de cavidade oral	72
Tabela 34 - Teste " <i>goodness-of-fit</i> " – modelo de regressão logística para tumores de cavidade oral	73
Tabela 35 - Teste " <i>goodness-of-fit</i> " – modelo de redes neurais para tumores de cavidade oral	73
Tabela 36 - Curvas de sobrevida em cinco anos para tumores de cavidade oral ...	74
Tabela 37 - Medidas de performance do modelo de CART para tumores de orofaringe	75
Tabela 38 - Medidas de R^2 e número de pacientes para cada subdivisão da amostra – modelo de CART para tumores de orofaringe	75
Tabela 39 - Curvas ROC do modelo de CART para tumores de orofaringe	76
Tabela 40 - Contribuição das variáveis preditivas para construção do modelo de BRT para tumores de orofaringe	77
Tabela 41 - Curvas ROC do modelo de BRT para tumores de orofaringe	77

Tabela 42 - Contribuição das variáveis preditivas para construção do modelo de Floresta Aleatória para tumores de orofaringe	78
Tabela 43 - Curvas ROC do modelo de Floresta Aleatória para tumores de orofaringe	79
Tabela 44 - Curvas ROC do modelo de redes neurais para tumores de orofaringe	79
Tabela 45 - Coeficientes estimados pela regressão logística: orofaringe	80
Tabela 46 - Razão de chances de cada variável para o modelo com tumores de orofaringe	81
Tabela 47 - Curvas ROC do modelo de regressão logística para tumores de orofaringe	81
Tabela 48 - Teste " <i>goodness-of-fit</i> " – modelo de regressão logística para tumores de orofaringe	82
Tabela 49 - Teste " <i>goodness-of-fit</i> " – modelo de redes neurais para tumores de orofaringe	82
Tabela 50 - Curvas de sobrevida em cinco anos para tumores de orofaringe	83
Tabela 51 - Medidas de R^2 e número de pacientes para cada subdivisão da amostra – modelo de CART para tumores de hipofaringe	84
Tabela 52 - Contribuição das variáveis no modelo de CART para hipofaringe	86
Tabela 53 - Curvas ROC do modelo de CART para tumores de hipofaringe	86
Tabela 54 - Contribuição das variáveis preditivas para construção do modelo de BRT para tumores de hipofaringe	87
Tabela 55 - Curvas ROC do modelo de BRT para tumores de hipofaringe	87
Tabela 56 - Contribuição das variáveis preditivas para construção do modelo de Floresta Aleatória para tumores de hipofaringe	88
Tabela 57 - Curvas ROC do modelo de Floresta Aleatória para tumores de hipofaringe	89
Tabela 58 - Curvas ROC do modelo de redes neurais para tumores de hipofaringe	89
Tabela 59 - Coeficientes estimados pela regressão logística: hipofaringe	90
Tabela 60 - Razão de chances de cada variável para o modelo com tumores de hipofaringe	91
Tabela 61 - Curvas ROC do modelo de regressão logística para tumores de hipofaringe	91

Tabela 62 - Teste " <i>goodness-of-fit</i> " – modelo de regressão logística para tumores de hipofaringe	91
Tabela 63 - Teste " <i>goodness-of-fit</i> " – modelo de redes neurais tumores em hipofaringe	92
Tabela 64 - Curvas de sobrevida em cinco anos para tumores de hipofaringe	92
Tabela 65 - Medidas de R^2 e número de pacientes para cada subdivisão da amostra - modelo de CART para tumores de laringe	93
Tabela 66 - Contribuição das variáveis no modelo de CART para tumores de laringe	95
Tabela 67 - Curvas ROC do modelo de CART para tumores de laringe	95
Tabela 68 - Contribuição das variáveis preditivas para construção do modelo de BRT para tumores de laringe	96
Tabela 69 - Curvas ROC do modelo de BRT para tumores de laringe	96
Tabela 70 - Contribuição das variáveis preditivas para construção do modelo de Floresta Aleatória para tumores de laringe	97
Tabela 71 - Curvas ROC do modelo de Floresta Aleatória para tumores de laringe	98
Tabela 72 - Curvas ROC do modelo de redes neurais para tumores de laringe	98
Tabela 73 - Coeficientes estimados pela regressão logística: laringe	99
Tabela 74 - Razão de chances de cada variável para o modelo com tumores de laringe	100
Tabela 75 - Curvas ROC do modelo de regressão logística para tumores de laringe	100
Tabela 76 - Teste " <i>goodness-of-fit</i> " – modelo de regressão logística para tumores de laringe	101
Tabela 77 - Teste " <i>goodness-of-fit</i> " – modelo de redes neurais para tumores de laringe	101
Tabela 78 - Curvas de sobrevida em cinco anos para tumores de laringe	102
Tabela 79 - Comparação dos modelos de regressão logística após a inclusão do HPV como variável independente	103
Tabela 80 - Parâmetros obtidos quando o HPV foi incluído no modelo de regressão para todos os sítios	103

Tabela 81 - Comparação dos modelos de regressão logística após a inclusão do HPV como variável independente na subamostras de teste	104
Tabela 82 - Parâmetros obtidos quando o HPV foi incluído no modelo de regressão para tumores de orofaringe	104
Tabela A.1 - Distribuição dos pacientes por centro	122
Tabela A.3 – Distribuição dos sociodemográfica	122
Tabela A.3 – Distribuição dos sintomas	123
Tabela A.4 – Distribuição de acordo com idade (anos)	123
Tabela A.5 – Distribuição de acordo com a perda ponderal (%) - apenas pacientes incluídos nas análises	124
Tabela A.6 – Distribuição de acordo com a quantidade de tempo que houve a perda ponderal (em meses) – apenas pacientes incluídos nas análises	125
Tabela A.7- Distribuição de acordo com o índice de massa corpórea (IMC) - apenas pacientes incluídos nas análises	125
Tabela B.1 – Dados obtidos após construção do modelo baseado em redes neurais para tumores todos os sítios	127
Tabela B.2 – Regressão logística para tumores de todos os sítios	127
Tabela B.3 – Dados obtidos após construção do modelo baseado em redes neurais para tumores de cavidade oral	127
Tabela B.4 – Regressão logística para tumores de cavidade oral	127
Tabela B.5 – Dados obtidos após construção do modelo baseado em redes neurais para tumores de orofaringe	128
Tabela B.6 – Regressão logística para tumores de orofaringe	128
Tabela B.7 – Dados obtidos após construção do modelo baseado em redes neurais para tumores de hipofaringe	128
Tabela B.8 – Regressão logística para tumores de hipofaringe	128
Tabela B.9 – Dados obtidos após construção do modelo baseado em redes neurais para tumores de laringe	129
Tabela B.10 – Regressão logística para tumores de laringe	129

RESUMO

Rios LFB. *Modelos de predição de risco de morte para pacientes com carcinoma epidermoide de cabeça e pescoço* [tese]. São Paulo: Faculdade de Medicina, Universidade de São Paulo; 2020.

INTRODUÇÃO: O carcinoma epidermoide de cabeça e pescoço (CECP) é uma das neoplasias malignas mais prevalentes no mundo. Assim como para diversas neoplasias, o Sistema TNM é a principal ferramenta para predição de sobrevida global em cinco anos e utiliza informações anatômicas, histopatológicas, funcionais e de infecção pelo Papilomavírus Humano (HPV). Contudo, a eficácia do TNM é questionada por categorizar pacientes em grupos de risco heterogêneos, além de não poder ser aplicado para predição de risco individualizado de morte. **OBJETIVO:** Criar uma ferramenta preditiva de mortalidade em cinco anos para pacientes com CECP para ser aplicada previamente ao tratamento, que fosse superior ao TNM, bem como avaliar a contribuição de dados sobre infecção pelo HPV nesta predição. **MÉTODOS:** estudo retrospectivo de dados prospectivamente coletados pelo grupo de pesquisa GENCAPO entre julho de 2000 e agosto de 2011. Todos os pacientes possuíam diagnóstico de carcinoma epidermoide da cavidade oral, orofaringe, hipofaringe e laringe, sem tratamento prévio e com ao menos cinco anos de seguimento. Os dados foram minerados utilizando Árvores de Classificação e Regressão, Florestas Aleatórias e *Boosted and Regression Trees*. Um modelo de regressão logística (RL) e outro com Redes Neurais (RN) foram criados para todos os sítios e para cada um dos sítios isoladamente. Métodos de validação interna, de discriminação por área sob a curva ROC (ASC) e calibração (*Goodness of Fit*) foram realizados. Amostras de tumor parafinadas de 692 pacientes foram testadas para o DNA do HPV por reação em cadeia de polimerase (PCR). O incremento preditivo do HPV foi analisado pelo coeficiente de determinação e pela ASC. **RESULTADOS:** A análise contou com 1811 pacientes e a taxa de positividade para o HPV foi de 8,7% para todos os sítios e 11,8% para orofaringe. Após a mineração de dados, 28 variáveis preditivas foram selecionadas. O modelo de RL para todos os sítios selecionou como variáveis significativas o volume clínico, sexo, idade e linfonodomegalia, com $ASC=0,77$. Já o modelo de RN obteve uma $ASC=0,76$. Os modelos de RL para cavidade oral e orofaringe selecionaram as mesmas variáveis que o modelo de todos os sítios, já o modelo de laringe excluiu sexo como variável, enquanto o modelo de hipofaringe selecionou apenas idade e linfonodomegalia. As medidas de ASC para cavidade oral foram: 0,78 (RL) e 0,74 (RN); para orofaringe de 0,72 (RL) e 0,65 (RN); para hipofaringe de 0,45 (RL) e 0,58 (RN); e para laringe de 0,71 (RL) e 0,76 (RN). Todos os modelos, exceto os criados para hipofaringe, apresentaram boa calibração. Comparando as curvas de sobrevida dos pacientes categorizados pelo risco calculado em quatro quartis ($Q1 \leq 25\%$, $25 < Q2 \leq 50\%$, $50 < Q3 \leq 75\%$ e $Q4 > 75\%$), com as curvas geradas sistema TNM, as primeiras apresentaram visualmente melhor separação entre as diferentes categorias de risco. Não houve incremento preditivo ao adicionar o HPV a nenhum dos modelos. **CONCLUSÃO:** Foi possível criar modelos preditivos com boa performance, a partir de informações facilmente coletáveis e superior ao TNM para predição de morte em cinco anos para pacientes com CECP. Em nossa população, não houve incremento preditivo ao incluir nos modelos o status HPV.

Descritores: Neoplasias de cabeça e pescoço; Mineração de dados; Predição; Modelos logísticos; Redes neurais; Árvores de decisões; Análise de sobrevida; Prognóstico.

ABSTRACT

Rios LFB. *Prediction models of death for patients with head and neck squamous cell carcinoma* [thesis]. Sao Paulo: “Faculdade de Medicina, Universidade de São Paulo”; 2020.

INTRODUCTION: Head and neck squamous cell carcinoma (HNSCC) is among the most prevalent neoplasms in the world. Likewise in other malignancies, the TNM System is the main tool for the prediction of 5-year overall survival and is based on anatomical, histological, and clinical information, as well as data on Human Papillomavirus (HPV) infection. However, the efficacy of this system is criticized by joining patients into heterogeneous categories, as well as for not being able to predict individualized probabilities of death. **OBJECTIVES:** To create a 5-year mortality prediction model to be applied at the first presentation, that performs better than TNM, as well as to evaluate the incremental contribution of HPV status to this model. **METHODS:** a retrospective study of data prospectively collected between July 2000 and August 2011 by the GENCAPO study group. Data were mined using Classification and Regression Trees, Random Forests, and Boosted and Regression Trees. Logistic Regression (LR) and Artificial Neural Network (ANN) models were created for all tumor sites and for each site separately. Methods of internal validation, discrimination - by the area under the ROC curve (AUC) - and calibration - Goodness of Fit test - were performed. Paraffin-embedded tumor samples from 692 patients were tested for HPV DNA by polymerase chain reactions (PCR). The contribution of HPV on model performance was analyzed by the coefficient of determination and the AUC. **RESULTS:** 1811 patients were analyzed, and HPV was positive in 8,7% of all tumors and in 11,8% of oropharyngeal tumors. After data mining the predictive variables, 28 were selected. The LR model for all sites selected tumor volume, sex, age, and lymph node enlargement as significant variables, with an AUC=0.77, while the ANN model obtained an AUC=0.76. The LR models for oral cavity and oropharynx tumors selected the same variables as the model for all sites, the LR model for larynx excluded sex as a significant variable, while the LR model for hypopharynx select only age and lymph node enlargement. The AUC for tumors of the oral cavity were 0.78 (LR) and 0.74 (ANN); for oropharynx were 0.72 (LR) and 0.65 (ANN); for hypopharynx were 0.45 (LR) and 0.58 (ANN); and for larynx were 0.71 (LR) and 0.76 (ANN). All models, except for the hypopharynx, presented good calibration. The survival curves according to the quartiles of predicted risk of death ($Q1 \leq 25\%$, $25 < Q2 \leq 50\%$, $50 < Q3 \leq 75\%$ e $Q4 > 75\%$) and according to the TNM stage were compared. The quartile curves presented a better visual separation than the TNM stage curves. There was no prediction increment by adding HPV status to any of the models. **CONCLUSION:** It was possible to develop models with good performance to predict 5-year survival, with easy to retrieve information, that were superior to the TNM System. No prediction increment was obtained by adding HPV status to the models in our study population.

Descriptors: Head and Neck Neoplasms; Data mining; Prediction; Logistic Models; Neural Networks; Decision Trees; Survival Analyses; Prognosis.

INTRODUÇÃO

1 – INTRODUÇÃO

O câncer de cabeça e pescoço tem atualmente uma incidência anual de aproximadamente 835.000 e uma mortalidade de 430.000 casos por ano(1). No Brasil, o Instituto Nacional do Câncer (INCA) estimou que em 2018 cerca de 23 pessoas a cada 100.000 habitantes terão câncer do trato aerodigestório alto(2). Atualmente representa o quinto sítio mais prevalente de câncer e o sexto em mortalidade relacionada à doença. Incluem nesta classificação de tumores do trato aerodigestório alto: cavidade nasal, seios paranasais, nasofaringe, orofaringe, cavidade oral, laringe e hipofaringe. Histologicamente pode-se subdividi-los em: sarcomas, melanomas, linfomas e carcinomas. O carcinoma epidermoide de cabeça e pescoço (CECP) constitui aproximadamente 85% destas neoplasias, possuindo comportamento muito variável(1).

O sistema TNM é a principal ferramenta para prever sobrevida nos pacientes acometidos com CECP. A última versão é a oitava e foi desenvolvida pela *Union for International Cancer Control* (UICC). Consiste em avaliar diversas características do tumor primário, dos linfonodos regionais e a presença de metástases à distância(3, 4), para posterior classificação do paciente em quatro grandes estágios. Sua principal crítica é o agrupamento de amostras muito heterogêneas de pacientes, de modo que pacientes com diferentes prognósticos sejam classificados em um mesmo grupo(5). Ademais, a simples classificação em categoria de risco nem sempre se traduz em uma aferição do risco individual de morte para determinado paciente.

Especialistas ao longo dos últimos anos estão questionando o TNM por não contemplar fatores de grande importância para prever o comportamento do CECP como a idade(6), presença de comorbidades(7, 8) e certas características tumorais(9). Na sua última edição, novos parâmetros já foram adicionados visando melhorar a precisão da classificação em estágios de pacientes portadores de CECP(4).

Além disso, já foi demonstrado que estratégias como as redes neurais artificiais são significativamente mais precisas do que o sistema de estadiamento TNM e que novas ferramentas prognósticas deveriam ser criadas para as mais distintas doenças oncológicas(10). Diversas propostas de desenvolvimento de ferramentas prognósticas mais precisas que o TNM já foram publicadas. Tais estudos concluem que é possível melhorar a predição de sobrevida global baseada na análise de dados clínicos.

Sabe-se ainda que o comportamento do CECP é influenciado por inúmeros fatores (genéticos, demográficos, culturais), o que acarreta dificuldade de aplicação de modelos desenvolvidos com dados de pacientes de contexto diferentes. Deste modo, a criação de um modelo nacional ajustado para a nossa população tornará a predição mais precisa para nossos pacientes (11).

Com o apoio da FAPESP e do Instituto Ludwig, em 2001 foi criado um grupo multicêntrico e multidisciplinar para o estudo do genoma do câncer de cabeça e pescoço. Posteriormente intitulado como GENCAPO, seu objetivo é não apenas o aprofundamento no conhecimento científico nas diferentes esferas da carcinogênese, como também o benefício prático e direto aos pacientes portadores de CECP(12).

Baseado neste banco de dados, o objetivo deste estudo é desenvolver modelos de predição mais precisos que o TNM ao estimar sobrevida para os pacientes com carcinomas epidermóides de cabeça e pescoço no momento de sua apresentação inicial, que seja aplicável para a população brasileira.

OBJETIVOS

2 – OBJETIVOS

2.1 - Identificar outros parâmetros não contemplados pelo TNM, que sejam significativos na predição de risco de morte em cinco anos para pacientes com carcinoma epidermoide de cabeça e pescoço.

2.2 - Criar um modelo de predição mais preciso que o TNM, a ser usado no momento da apresentação inicial.

2.3 - Avaliar o incremento do status HPV do tumor no valor de predição de óbito do paciente.

REVISÃO DA LITERATURA

3 – REVISÃO DA LITERATURA

3.1 – Epidemiologia – Carcinoma epidermoide no Brasil e no mundo

O perfil epidemiológico do carcinoma epidermoide do trato aerodigestório alto vem apresentando grandes mudanças nos últimos anos(13). Historicamente é um tipo de neoplasia maligna que sempre esteve fortemente associado ao tabagismo e etilismo, sendo mais comum em homens a partir da sétima década de vida(13). Graças às políticas antitabágicas, o consumo de tabaco reduziu bastante, principalmente em regiões mais desenvolvidas(14). Desta forma houve uma redução gradativa da incidência deste tipo de doença ao longo dos dois últimos decênios, embora menos do que seria esperado, graças ao aparecimento do HPV como um fator de risco importante(15). Tal condição, além de outros fatores ambientais ainda desconhecidos, levou em algumas regiões pelo mundo, à mudança do perfil epidemiológico do CECP, com o aumento considerável de novos casos em pacientes mais jovens e mulheres, especialmente de tumores em subsítios da orofaringe, sobretudo loja amigdaliana(16).

Dados do GLOBOCAN mostram que atualmente 835.000 novos casos de CECP do trato aerodigestório alto são diagnosticados anualmente no mundo(1), com uma mortalidade em torno de 430.000 casos por ano(1, 17, 18). O Brasil acompanha o perfil epidemiológico dos países em desenvolvimento. Dados do INCA (Tabela 1) mostram que a estimativa para 2018 era diagnosticar cerca de 17.590 homens e 4.780 mulheres com neoplasias malignas da cavidade oral, faringe e laringe(2). O câncer de cavidade oral é o quinto mais frequente em homens e o décimo segundo mais frequente em mulheres, ao passo que o câncer de laringe é o oitavo mais frequente em homens e o décimo sexto mais frequente em mulheres(2). Ao analisar os dados de forma regionalizada e considerando todos os sítios de neoplasia maligna, exceto pele não melanoma, conclui-se que a região sudeste é onde ocorre a maior incidência das neoplasias de cavidade oral e laringe, ao passo que a região nordeste é onde há a menor incidência proporcional(4).

Tabela 1 – Estimativa do número de casos novos segundo sexo, por região.

Sítio	Sexo	Norte	Nordeste	Sudeste	Sul	Centro Oeste
Laringe	M	250 (3,4%)	1430 (2,7%)	2720 (2,8%)	1540 (4,8%)	420 (2,9%)
	F	80 (0,9%)	310 (0,6%)	300 (0,3%)	210 (0,5%)	90 (0,7%)
	Total	330 (4,3%)	1740 (3,3%)	3020 (3,1%)	1750 (5,3%)	510 (3,6%)
Cavidade Oral	M	290 (3,4%)	1880 (3,6%)	6030 (6,3%)	2260 (4,2%)	680 (4,7%)
	F	160 (1,9%)	1190 (2,2%)	2300 (2,3%)	490 (1,1%)	210 (1,5%)
	Total	450 (5,3%)	3070 (5,8%)	8330 (8,6%)	2750 (5,3%)	890 (6,2%)

M: masculino; F: feminino. Entre parênteses encontra-se a proporção em relação ao número absoluto de novos casos de câncer (exceto pele não melanoma). FONTE: INCA, 2018

3.2 –TNM

A previsão de desfechos é extremamente comum em oncologia, pois inúmeras decisões a respeito do tratamento, expectativa do paciente e de familiares, além de decisões pessoais, dependem de seus resultados esperados.

Podem-se aplicar ferramentas preditivas em diversas situações em oncologia, como: previsão de risco aumentado de um câncer que justifique a indicação de rastreamento, predição de resposta a determinado tratamento, prever a ocorrência de metástases ou óbito em um certo intervalo de tempo etc.

Historicamente o primeiro modelo visando predição em oncologia foi criado por Pierre Denoix entre os anos de 1943 e 1952, sendo chamado de sistema TNM(19). Alguns anos depois, as primeiras orientações relacionadas ao estadiamento oncológico (mama e laringe) foram publicadas(19). Nos decênios seguintes, ocorreram sucessivas melhorias no sistema TNM, como a ampliação para outros sítios, assim como a incorporação de fatores prognósticos não anatômicos. Tais melhorias são desafiadoras para o sistema TNM, pois apesar de teoricamente tornarem a predição mais precisa, podem levar a uma complexidade que o torne pouco prático.

O sistema TNM é uma ferramenta amplamente utilizada pelos profissionais de saúde e pela comunidade científica, por ser um modelo relativamente simples no qual parâmetros anatômicos e não anatômicos relacionados ao tumor são avaliados, tabulados e agrupados. Não há necessidade de cálculos ou *softwares* para classificar um doente. A

simplicidade era o objetivo no TNM, já que a ferramenta foi idealizada em uma época anterior à informatização. Atualmente, o acesso a dispositivos móveis capazes de cálculos complexos de maneira rápida e precisa também é universal.

Na última edição do TNM para CECP, novos parâmetros foram adicionados: espessura tumoral para tumores em cavidade oral, presença de extravasamento extracapsular no linfonodo e pesquisa do p16 para aqueles em orofaringe(4).

O TNM adota a estratégia de dividir os tumores do trato aerodigestório alto de acordo com o sítio (Figura 1). Para cavidade oral este modelo contempla as três dimensões tumorais e se há invasão de ossos da face, pele, espaço mastigatório, processos pterigoideos, envolvimento da artéria carótida interna ou base do crânio. Em relação aos tumores da faringe, é proposto uma divisão de acordo com cada subsítio: orofaringe, nasofaringe ou hipofaringe. De maneira geral, para classificar os tumores da faringe é fundamental saber o tamanho do maior diâmetro, a localização e a extensão para estruturas extrafaríngeas. Na última edição foi incluída a pesquisa do HPV para tumores de orofaringe, seja por PCR, seja por imunohistoquímica com p16. Já para os tumores laríngeos, o sistema TNM adota uma avaliação funcional associada, pois é necessário saber se as pregas vocais estão móveis, paréticas ou fixas, além também de considerar extensão a outros subsítios laríngeos ou extralaríngeos. No caso das metástases linfonodais regionais, o TNM contempla o número de linfonodos, a lateralidade, o tamanho e, mais recentemente, a presença de sinais de extravasamento extracapsular(20).

Nos últimos anos, o número de pesquisadores que defendem a criação de algoritmos mais elaborados para predição em câncer vem crescendo. Modelos antigos baseados nos mesmos critérios utilizados pelo sistema TNM, vêm sendo gradativamente substituídos por ferramentas estatísticas mais robustas(21). Com o aparecimento de descobertas relacionadas ao comportamento tumoral, novas informações são coletadas e disponibilizadas na área de pesquisa médica.

Figura 1 - Parâmetros utilizados para definir o estágio do sítio primário no TNM (cavidade oral, faringe e laringe)

3D: três dimensões; HPV: papilomavírus humano; PPVv: pregas vocais

3.2.1 – Incorporação do HPV ao TNM

Na última década houve um aumento na incidência cumulativa de CECP da orofaringe, particularmente em homens brancos de meia idade com etilismo moderado e exposição limitada ou ausente ao fumo. Epidemiologicamente, isso se deve ao aumento da incidência de cânceres HPV de alto risco(22, 23).

Em 2015 um estudo canadense demonstrou a importância em incluir o HPV no estadiamento do CECP para orofaringe, pois pacientes com tumores relacionados ao HPV tinham um prognóstico melhor do que aqueles não relacionados(24). No ano seguinte um estudo de coorte multicêntrico examinou retrospectivamente os dados de 1907 pacientes, elaborou e validou um novo método para classificar estes pacientes, sendo proposta uma mudança no TNM 8ª edição(25). Outros estudos validaram retrospectivamente a última edição do TNM na Austrália(26), Alemanha(27) e no Japão(28).

O sistema de estadiamento TNM da 8ª edição incorporou o HPV no prognóstico dos pacientes com CECP de orofaringe portadores de HPV e estudos mostraram que tal mudança parece estratificar melhor estes pacientes(29).

3.2.2 - Fatores prognósticos não contemplados pelo TNM em CECP

Fatores preditivos de sobrevida, como a idade, tabagismo, etilismo, raça, escolaridade, aspectos sociais, presença de comorbidades, performance *status* e nutrição tem sido propostos para incorporação ao TNM (6, 8, 30). Estes estudos deixam claro que

tais fatores impactam na sobrevida, mas não há nenhuma grande análise abrangente a respeito de como se comportam tantos fatores em um único modelo preditivo.

3.2.2.1 – Tabagismo

Em 1957 Wynder demonstrou através de um estudo de caso-controle, a associação entre o tabagismo e o câncer oral(31). No ano seguinte, Hammond apresentou durante o congresso da *American Medical Association*, em Nova York, um estudo de coorte envolvendo mais de 180.000 pacientes demonstrando como as pessoas que fumavam tinham um risco maior de desenvolver CECP do que aquelas que nunca fumaram(32).

O tabagismo é um fator de risco independente bem consolidado no que se refere ao desenvolvimento do carcinoma epidermoide do trato aerodigestório alto. Apesar do consumo de tabaco reduzir ao longo dos últimos anos, este fator ainda é, globalmente, o mais importante fator de risco para CECP, principalmente em áreas menos desenvolvidas. A *International Agency for Research on Cancer* (IARC) classifica o tabagismo como Grupo 1: carcinogênico para o ser humano(33). Tal fato ocorre devido a presença de vários carcinógenos que levam a alterações genéticas importantes seguidas do desenvolvimento tumoral. Contudo, o tabagismo também parece ser um fator prognóstico.

Em 2004 Pytynia *et al.* publicaram um estudo no qual foi realizado uma análise de sobrevida entre pacientes com diagnóstico de CECP, sendo 50 tabagistas e 50 não tabagistas(34). Concluíram que o grupo dos tabagistas possui um risco de morte significativamente maior quando comparado com o outro grupo ($RR=3,50$; 95% IC, 1,14-10,77; $P=0,029$). Vários outros estudos encontraram respostas semelhantes, mantendo a conclusão de que os pacientes tabagistas possuem uma sobrevida menor que aqueles não tabagistas. A sobrevida global costuma ser afetada pelo tabagismo, provavelmente por causa da relação entre este hábito com complicações perioperatórias e com comorbidades, como as cardiovasculares e doenças pulmonares crônicas(35).

3.2.2.2 – Raça

Inúmeros estudos mostram uma importante diferença na sobrevida entre negros e brancos diagnosticados com CECp(30). Outra informação relevante é que não há grandes diferenças na incidência do câncer de cabeça e pescoço entre negros e brancos. Dados da *American Cancer Society* mostram que a mortalidade do câncer de cavidade oral e orofaringe na população branca era de 3,7/100.000, enquanto que na população negra era de 6,5/100.000 entre os anos de 2002-2006(36). Molina *et al* encontraram achados semelhantes e dados norte-americanos revelaram que a sobrevida global em cinco anos para câncer de cabeça e pescoço era de 59% em brancos e de 39,5% em negros(37). Mesmo que as diferenças de sobrevida entre negros e brancos possam ser atribuídas em parte a uma combinação de elementos (socioeconômicos, acesso a saúde, etc.), estudos mostram que a disparidade permanece mesmo quando ajustadas a tais fatores(36, 38).

No Brasil, a miscigenação da população torna a análise de dados raciais de complexa execução e tal dificuldade é ainda intensificada pelo fato de que o simples registro da cor autodeclarada pelo paciente muitas vezes não é correlacionado com a herança genética(39, 40).

3.2.2.3 – Idade

Muitos estudos mostram que a idade é um excelente fator preditivo de mortalidade, pela presença de comorbidades ou pelo tratamento indicado às pessoas mais idosas(41). Udeabor *et al* conduziram uma análise retrospectiva para avaliar a sobrevida global de pacientes com diagnóstico de câncer de cabeça e pescoço comparando pacientes com menos e mais de 40 anos de idade. Eles concluíram que a sobrevida global para o primeiro grupo era melhor que no segundo grupo e que os pacientes com comorbidades mais importantes são apresentam sobrevida inferior(42). Estudo do Gencapo não demonstrou diferença em sobrevida global de pacientes maiores e menores que 40 anos, mesmo quando estratificados por subsítios tumorais.

3.3 - Predição em outras áreas da oncologia

Diversos modelos em câncer além do TNM já foram desenvolvidos,

especialmente em câncer de mama e próstata, dada as suas altas prevalências. Merece destaque a evolução que modelos de predição em câncer de mama vêm apresentando. Ferramentas que antes incorporavam apenas informações clínicas, como idade cronológica da paciente, da menarca, do nascimento do primeiro filho, etnia, história familiar dentre outras, foram sendo atualizados com informações acerca de mutações genéticas, dentre elas o *status* BRCA1/2(43-46). No caso do câncer de próstata os inúmeros modelos desenvolvidos para predição de risco possuem diversos objetivos como: indicar rastreamento e biópsia(47, 48), resposta terapêutica (49-51), disseminação metastática e mortalidade(52).

3.4 - Novos conceitos

Em um mundo digitalizado, a quantidade de informação aumenta exponencialmente (53). Por outro lado, a heterogeneidade dos dados dificulta análises mais amplas que somente serão viáveis com a identificação de padrões relevantes (54). O aprendizado de máquina é uma das estratégias mais empregadas para lidar com essa complexidade ao prever um resultado(55). É um processo no qual a modelagem é alcançada através do aprendizado baseado na experiência. Essa ferramenta faz parte do campo da inteligência artificial e é amplamente usada em diferentes áreas da ciência (56). O aprendizado profundo é uma ampla classe de técnicas de aprendizado de máquina para extrair abstrações de alto nível de uma grande quantidade de dados brutos e heterogêneos (53, 57) (Tabela 2).

3.5 - Desenvolvimento de modelos preditivos em oncologia

Ao longo do tempo, vários modelos preditivos para HNSCC foram desenvolvidos (58-61). A maioria destes modelos utilizou bancos de dados robustos com variáveis como o sítio tumoral, idade, extensão do tumor, metástases linfonodais e sua extensão extracapsular, escolaridade, tabagismo, contagem de células sanguíneas, entre outras. Algumas limitações referem-se à complexidade da doença ou à sua natureza retrospectiva. No entanto, vários estudos incorrem em erros desnecessários devido ao uso inadequado de técnicas de pesquisa e de determinados métodos estatísticos. Listaremos a

seguir os principais requisitos necessários para o adequado desenvolvimento e aplicação de modelos preditivos, bem como as principais ferramentas disponíveis.

Tabela 2 – Definição de termos comumente usados na predição moderna.

Termo em inglês conhecidos globalmente (tradução)	Conceito
<i>Big data</i>	um campo que maneja maneiras de analisar sistematicamente, extrair informações ou lidar com conjuntos de dados muito grandes ou complexos para serem manipulados pelo software tradicional de processamento de dados
<i>Machine Learning</i> (máquina de aprendizado)	estudo científico de algoritmos e modelos estatísticos que os sistemas de computador usam para executar uma tarefa específica sem usar instruções explícitas, com base em padrões e inferência.
<i>Deep learning</i> (aprendizagem profunda)	uma ampla classe de técnicas de aprendizado de máquina para extrair abstrações de alto nível de uma grande quantidade de dados brutos e heterogêneos
<i>Artificial intelligence</i> (inteligência artificial)	capacidade do sistema de interpretar corretamente dados externos, aprender com esses dados e usá-los para atingir metas e tarefas específicas através da adaptação
<i>Data mining</i> (mineração de dados)	processo para explorar grande quantidade de dados, procurando padrões consistentes, associações e sequências temporais

3.5.1 - Coleta de Dados

A forma como os dados a serem analisados foram coletados é fundamental para a boa qualidade dos resultados em qualquer análise estatística. Em modelos de predição de risco o ideal é que os dados tenham sido obtidos através de uma coorte prospectiva. Dados oriundos de estudos randomizados são mais difíceis de serem generalizados devido aos critérios de inclusão, os quais costumam ser mais rigorosos. Já os dados colhidos por outros motivos como prontuários médicos, laudos de exames etc., costumam apresentar um período de acompanhamento satisfatório, mas podem apresentar dados incompletos, além de terem sido obtidos de maneira não sistematizada, com maior probabilidade de erros de aferição e possivelmente não sendo generalizáveis para outros contextos(62).

3.5.2 - Análise de consistência

Lidar com dados médicos envolve transferir dados coletados em um ambiente clínico para um servidor, verificar e limpar os dados para remover erros e verificar inconsistências(63). A análise de consistência dos dados é uma das etapas mais críticas na construção de um modelo preditivo. Erros de registro são comuns e podem ser reduzidos ou mesmo eliminados com uma revisão minuciosa. Registros duplos das mesmas informações (exemplo: data de nascimento e idade) podem ajudar a manter a precisão. Embora os sistemas eletrônicos de coleta de dados já tenham sido projetados para evitar erros, uma revisão crítica da consistência é indispensável. Uma análise cuidadosa dos valores discrepantes também pode ser valiosa.

3.5.3 - Exploração de dados

Técnicas que detectam tendências e padrões ocultos nos bancos de dados são necessárias para pesquisar grandes quantidades de dados (64). Essa análise prévia pode orientar para melhores maneiras de incorporar variáveis aos modelos de predição. Por exemplo, ao usar um método comum como um modelo de regressão, o pesquisador pode decidir transformar uma variável contínua em uma categórica, de acordo com a análise pregressa ou na forma como essa variável pode se comportar, encontrando, por exemplo, pontos de corte ideais ou uma relação de comportamento não linear de uma variável contínua.

3.5.4 - Seleção de variáveis

A grande quantidade de variáveis preditivas está se tornando cada vez mais comum. Se por um lado, isso é bom, pois pode propiciar a descoberta de preditores até então desconhecidos, por outro, pode ser um desafio para o desenvolvimento de modelos preditivos, especialmente se o tamanho da amostra não for tão grande. Mais precisamente, a frequência dos desfechos é o principal parâmetro para se prever o poder de detecção de um modelo. A exatidão, precisão e significância dos coeficientes estimados pelo método de risco proporcional se tornarão não confiáveis quando a taxa de eventos por variável for muito baixa. Harrell *et al.* sugere como regra geral aproximada que, em uma análise

de regressão, para que haja discriminação preditiva, para cada variável preditiva deve haver dez sujeitos que tiveram o evento como resultado na amostra de treinamento. Portanto, se o número de variáveis preditivas for maior que um décimo do número de indivíduos com o resultado, deve ser usada uma técnica de redução de dados, como componentes principais, agrupamento de variáveis ou resumo clínico derivado(65). Assim, se uma amostra for grande, porém com desfechos pouco frequentes, por exemplo morte, seu poder de predição pode ser limitado. Assim, observa-se o desfecho menos frequente (que pode ser morte ou sobrevida) e se prevê a capacidade de incorporação de variáveis do modelo(66).

3.5.4.1 - Categorização de variáveis

A categorização de variáveis contínuas é frequentemente usada pois cria faixas de corte que são mais facilmente utilizáveis na prática clínica. Essa estratégia facilita a análise e interpretação de dados contínuos, pois evita certas suposições sobre a natureza dos dados (por exemplo, linearidade). No entanto, essa categorização também pode gerar problemas, principalmente por reduzir o poder estatístico da amostra devido à perda de informações. Pacientes semelhantes podem ser categorizados em categorias distintas, de acordo com os pontos de corte escolhidos(67). Um maior número de categorias pode minimizar a perda de informações, portanto, a categorização de variáveis contínuas deve ser cuidadosamente avaliada antes da análise, levando em consideração os prós e os contras(68).

3.5.4.2 - Análise de componentes principais (PCA)

A riqueza de um modelo matemático está na variabilidade de seus dados. A análise de componentes principais (PCA) é um algoritmo matemático que reduz a quantidade de dados enquanto mantém a variação desejada no conjunto dos mesmos. Ele realiza a redução dos dados identificando direções, chamadas "componentes principais", ao longo das quais a variação nos dados é máxima. Cada sujeito pode ser representado por poucos números, que são resultados de equações compostas pelas inúmeras variáveis, e não mais por essas variáveis. Essa técnica também permite que padrões ocultos na

grande quantidade de dados sejam identificados após a redução no número de variáveis(69).

3.5.5 - Aditividade dos efeitos preditivos e interações entre as variáveis

Duas outras questões importantes na aplicação de modelos de regressão são a aditividade dos efeitos preditivos e a interação entre as variáveis. Quando dois fatores preditivos não interagem, é possível afirmar que cada variável tem um efeito aditivo na resposta, ou seja, um modelo de regressão possui efeitos aditivos se a função da resposta poderá ser descrita como a soma das funções das variáveis preditivas. Por exemplo, ao se estimar o peso de nascimento de um bebê em gramas(y) pode-se usar como variáveis preditivas a duração da gestação em semanas (x_1) e o fato da gestante ser tabagista ou não-tabagista (x_2). Neste caso o modelo preditivo irá fornecer um coeficiente (β_1) para a variável x_1 . Supomos que seja 150, ou seja, para cada semana de gestação presume-se que haja um ganho de 150g do peso ao nascer. O tabagismo (x_2) por sua vez, neste modelo, é a outra variável que influencia no peso ao nascer, mas caso também interfira na duração da gestação, existirá interação entre as variáveis e a presunção de aditividade será rompida.

Com relação à interação, o número de possíveis resultados cruzados geralmente é tão grande que os preditores para verificar a aditividade ou interação devem geralmente ser especificados antes de se testar os preditores, caso contrário, os erros de tipo I, ou seja, achados falsos positivos e a superposição serão problemas importantes. Utilizando o mesmo exemplo acima, se de fato houver interação entre o tabagismo e a duração da gestação e esta interação não for identificada antes da construção do modelo, a equação pode ser erroneamente calculada e sua habilidade preditiva comprometida.

Uma solução para estas questões é fazer testes de interação agrupados. Por exemplo, em um modelo com preditores de idade, sexo e carga tabágica, pode-se testar todas as interações de segunda ordem envolvendo idade, todas as interações envolvendo sexo e todas as que envolvem a carga tabágica. Um teste combinado de interações bidirecionais também é útil. Se um teste agrupado não atingir significância, pode ser um erro buscar resultados significativos através destas interações. Contudo, tem-se recomendado utilizar outros métodos estatísticos de mineração de dados para buscar

interações e relações não lineares, a fim de se identificar a melhor forma de inserir essas variáveis no modelo, mesmo que ainda se prefira o uso da regressão logística na sua construção final, dada sua facilidade de interpretação e abrangência de uso.

Durante a construção dos modelos preditivos, deve-se analisar se as variáveis de preditores contínuos apresentam relações lineares com o desfecho, pois caso não possuam, são necessários ajustes na forma em que se analisam estes preditores. Por exemplo, a relação entre a idade e a mortalidade por pneumonia não é linear, pois a pneumonia costuma apresentar maiores taxas de mortalidade nos extremos da idade. Não considerar a relação quadrática da idade em um modelo preditivo para mortalidade de pneumonia pode significar não conseguir identificar o poder preditivo desta variável, quando ele de fato existe.

3.5.6 - Dados faltantes

Dados faltantes constituem um problema inerente à natureza da pesquisa biológica. Sempre que possível, a exclusão de pacientes por dados faltantes deve ser evitada, pois além de reduzir o tamanho da amostra, pode levar a redução da acurácia da ferramenta, com possível geração de vieses, já que muitas vezes a ausência dos dados não ocorre de forma aleatória. Em regressão logística, um método frequentemente utilizado para desenvolvimento de modelos preditivos, apenas casos com dados completos para todas as variáveis poderão ser incluídos. Deste modo, dados faltantes em qualquer variável ameaçam a construção do modelo de forma relevante. Neste caso é importante buscar incessantemente dados faltantes e, no caso da impossibilidade de se obtê-los, utilizar técnicas de imputação de dados, dando preferência para imputação multivariada(70). Caso a quantidade de pacientes com uma determinada variável faltante seja considerável, existe a possibilidade de tal informação também ser de difícil obtenção na prática. Nestes casos, deve-se questionar a manutenção desta variável no modelo(71). É importante lembrar que dados faltantes de variáveis dependentes, ou seja, do desfecho, jamais devem ser imputados.

A proporção de dados ausentes está diretamente relacionada à qualidade das inferências estatísticas. No entanto, não há consenso sobre um valor de corte de uma porcentagem aceitável de dados faltantes em um conjunto de dados(72). Se sua proporção

for pequena (5% ou menos), a imputação simples poderá ser bastante razoável (73). Por outro lado, se a proporção de dados ausentes for superior a 10%, os resultados de análises estatísticas subsequentes poderão ser enviesados (74). Os mecanismos e padrões de dados faltantes têm maior impacto nos resultados da pesquisa do que a proporção de dados faltantes em si (75).

3.5.6.1 – Aleatoriedade de dados faltantes

Embora comum em vários estudos, assumir a aleatoriedade de dados faltantes é arriscado. É improvável que, para qualquer variável, os valores não faltantes constituam definitivamente uma amostra aleatória simples dos valores para todos os sujeitos do estudo(76, 77). A questão principal é se a "falta" está relacionada aos valores dos dados ou não.

3.5.6.2 - Recomendações e limitações da imputação de dados

Existem várias formas para manejar a falta de informação em um banco de dados. Métodos simples e frequentemente utilizados incluem a análise completa apenas dos dados disponíveis, que pode levar a um aumento enviesado da variância(76), e a imputação utilizando como critério a média geral dos dados acessíveis(78). O problema é que estes métodos levam a análises ineficientes e geralmente produzem estimativas de associação tendenciosas. Existem técnicas mais sofisticadas e recomendadas de imputação para lidar com dados faltantes, como a imputação multivariada.(76). Com essa técnica, os dados faltantes para um determinado indivíduo são imputados por um valor que é previsto usando as outras características disponíveis do mesmo, em modelos desenvolvidos a partir do comportamento previsto pelos dados completos de outros pacientes.

Uma análise de sensibilidade comparando diferentes estratégias para lidar com dados faltantes também pode fornecer informações sobre a importância de diferentes métodos de imputação e como os dados faltantes podem prejudicar o modelo se não forem tratados adequadamente.

3.5.7 - Ajustes para análises múltiplas

Na medida em que o número de testes aumenta, aumenta também a probabilidade de um erro do tipo I, ou seja, concluindo que uma diferença significativa está presente quando não está. Por exemplo, em um estudo com 20 análises e nenhuma relação verdadeira entre preditores e resultados, com $p \leq 0,05$ como nível crítico para todos os testes, pode ser demonstrado que a chance de pelo menos um teste ser estatisticamente significativo por acaso não é 5%, mas sim 64%(79). Inúmeros métodos são usados para ajustar os valores de p ao fazer vários testes estatísticos com o objetivo de evitar esse erro. A correção de Bonferroni é direta e amplamente utilizada (80), apesar de ser considerada muito rigorosa (81). Ele estima a taxa de erro pela fórmula α/T , onde α é o valor crítico escolhido de p e T o número de testes realizados. O resultado é aplicado a cada teste individual para manter o nível geral de significância em 0,05. Há um debate contínuo sobre esse método, especialmente em estudos exploratórios, onde pode ser muito rigoroso e comprometer importantes descobertas. Outros métodos menos conservadores, como taxas de falsas descobertas, entre outros, têm sido propostos para diminuir o risco de falha na detecção de associações importantes(81).

3.5.8 - Métodos estatísticos para modelos preditivos

Durante décadas, os métodos mais utilizados para construir modelos de predição foram regressões e redes neurais artificiais(82, 83), sendo o segundo o preferido quando se lidam com amostras grandes(84). Outros métodos estatísticos tradicionalmente empregados em predição e prognóstico são as análises discriminativas(85), abordagem Bayesiana (86) e regressões múltiplas(87). Estes modelos se apresentaram efetivos, com a ressalva de resolver problemas menos complexos. Mais recentemente, a evolução das redes neurais e outros métodos de reamostragem, como Floresta Aleatória e *Boosted Regression Trees* (BRT), também têm sido utilizados para melhor manejo de problemas complexos (88, 89).

3.5.8.1 - Regressões

Regressão é um dos métodos estatísticos mais tradicionalmente usados para

modelagem preditiva e mineração de dados. Existem mais de 12 tipos de algoritmos de regressão projetados para diferentes tipos de análise, incluindo os três mais utilizados: regressão linear, logística e Cox. Cada tipo tem sua importância e é usado para modelar o relacionamento entre uma variável dependente (desfecho) e uma ou mais variáveis independentes (preditores).

Cada técnica de regressão tem suas próprias presunções implícitas, que devem ser atendidas antes da execução da análise. Essas técnicas diferem em termos do tipo de variáveis dependentes e independentes e suas distribuições.

A regressão linear é usada quando a variável dependente é contínua e sua relação com as variáveis independentes é assumida como linear. A regressão de Cox(90) é usada para investigar a associação entre o tempo de sobrevivência dos pacientes e uma ou mais variáveis preditoras, especialmente úteis quando a perda de acompanhamento é uma questão importante.

Em situações nas quais se busca estimar um desfecho binário (vivo ou morto, falha ou sucesso, sim ou não) baseado em uma ou mais variáveis contínuas ou categóricas, pode-se utilizar a regressão logística(91). Este recurso estatístico possui capacidade de avaliar as razões de chances individualizadas e, se inseridos os termos apropriadamente no modelo, a interação entre as variáveis. Evita-se assim analisar as variáveis preditivas de maneira isolada, pois desta forma a covariância entre elas seria ignorada e isto poderia levar a achados não verdadeiros(92).

Inicialmente o programa estatístico calcula o risco de base de ocorrer ou não o desfecho sem usar qualquer preditor. Esta análise inicial fornece a constante (também conhecida como a intercepção). Em seguida, as variáveis preditivas escolhidas são inseridas no modelo e um coeficiente de regressão (conhecido também como "beta") e um valor de significância "p" para cada uma destas variáveis é calculado. O valor "p" indica se a variável particular contribui significativamente para a ocorrência do resultado ou não. Esses resultados também podem ser expressos como uma equação, que inclui a constante e o coeficiente de regressão para cada variável, que foi considerada significativa (geralmente usando um valor de $p < 0,05$ como limiar). A equação fornece um modelo que pode ser usado para prever a probabilidade de um evento acontecer para um indivíduo em particular, dado seu perfil de fatores preditores(91) (Figura 2).

Figura 2 - Equação modelo na regressão logística.

Equação de regressão logística para um sujeito (i) onde y é o *log-odds* do resultado, β_0 é a interceptação, β é o coeficiente de cada variável "x". A partir do *log-odds*, as chances e probabilidades podem ser calculadas.

3.5.8.1.1 - Seleção de variáveis em modelos de regressão

Em modelos preditivos, a escolha das variáveis é fundamental para que se consiga prever o desfecho com menor possibilidade de erro. Apesar de ser frequentemente observado em diversos estudos, esta seleção não deve ser oriunda de análises univariadas, pois levaria a conclusões possivelmente errôneas(66, 93). Além disso, todas as variáveis precisam estar associadas ao desfecho, mas tal correlação não precisa ser necessariamente causal(71). Para a escolha do número dessas variáveis, é importante que esteja adequado à quantidade do desfecho menos frequente encontrado na amostra, para que os modelos resultantes não se tornem instáveis e imprecisos(65).

Com o propósito de identificar quais variáveis irão compor o modelo preditivo, duas estratégias são bem difundidas na literatura: modelo cheio e estratégia de seleção dos preditores. No primeiro, todas as variáveis selecionadas são incluídas nas análises multivariadas e nenhum método de seleção de preditores é utilizado. Aqueles que defendem este método alegam que tal conduta reduz o viés de seleção de variáveis preditivas, bem como o sobreajuste do modelo(66, 94). O modelo cheio, entretanto, não é geralmente fácil de ser definido, pois requer conhecimento prévio a respeito das variáveis preditivas candidatas. Deve-se evitar esta estratégia quando o número de eventos é limitado ou quando existem muitas variáveis candidatas(94).

A outra estratégia é fazer a seleção dos preditores através de análises multivariadas com métodos para maximizar medidas de performance dos modelos, como o máximo R^2 ajustado, *Akaike information criterion* (AIC)(95), *Bayesian information criterion* (BIC), dentre outros(96). As variáveis candidatas que não contribuem no modelo são removidas. Na seleção passo a passo (*stepwise selection*) é realizada uma

série de testes removendo, adicionando ou mantendo cada variável no modelo baseado em níveis de significância predefinidos. Este processo pode ser realizado em diferentes “direções”. Pode-se realizar a eliminação retrógrada, na qual a modelagem se inicia com todas as variáveis candidatas incluídas no modelo e são sequencialmente eliminadas, ou, de maneira oposta, na seleção anterógrada, inicialmente sem variáveis e estas são incluídas no modelo uma a uma, conforme o comportamento da medida de performance escolhida. Esta última estratégia é menos utilizada quando comparada com a eliminação retrógrada, pois não demonstra as interações de todas as variáveis entre si, contudo, é especialmente útil quando muitas variáveis candidatas são propostas. Existe ainda a tática que combina as duas supracitadas. Através desse último método pode-se adicionar ou remover as variáveis em inúmeras etapas, com vantagens quando existem inúmeras variáveis potenciais, porém com a possibilidade de entrada de variáveis produto de ruído entre as que já estão no modelo. Experiência clínica e opinião não são permitidas na escolha inicial das variáveis(96), porém podem ser usadas posteriormente para forçar variáveis sabidamente relevantes ao modelo posteriormente proposto.

3.5.8.2 - Redes Neurais

O termo "redes neurais" refere-se a um conjunto de unidades de processamento (nós) que simulam neurônios e são interconectados por meio de "pesos" (análogos às conexões sinápticas no sistema nervoso) para permitir que os sinais atravessem a rede paralelamente e em série. Os nós são elementos computacionais diretos e baseiam-se na observação de que um neurônio se comporta como um interruptor: quando uma certa quantidade de neurotransmissor se acumula no corpo da célula, um potencial de ação é gerado. Isso foi modelado matematicamente como uma soma ponderada de todos os sinais de entrada para um nó, que é comparado com um limite que, quando excedido, leva ao disparo do nó, caso contrário, permanece inativo. O poder computacional em uma rede neural não deriva da complexidade de cada unidade de processamento (como em um computador convencional), mas da densidade e complexidade das interconexões. Ao contrário dos computadores tradicionais, a memória em uma rede neural é distribuída por toda a estrutura (em pesos) e é modificada pela experiência(97).

Nesse modelo matemático, o peso das variáveis é equivalente aos coeficientes beta nas regressões logísticas, e o fluxo de informações através dos nós é unidirecional. Este modelo pode aprender relações matemáticas entre variáveis preditivas e resultados. Possui métodos de validação interna e geralmente cria modelos com excelente precisão(98). Muitos consideram que, ao analisar dados com muitos parâmetros de entrada, as redes neurais são essenciais, pois nenhum método baseado em regressão é adequado (99).

Em um estudo publicado em 1997, Burke *et al*(10) concluíram que as redes neurais artificiais eram significativamente mais precisas do que o sistema de estadiamento TNM e que novas ferramentas prognósticas deveriam ser criadas nas mais diversas doenças oncológicas. Jones *et al.* concluíram que as redes neurais melhoram a previsão da sobrevida observada em pacientes com carcinoma epidermoide da laringe (100).

3.5.8.3 - Métodos baseados em árvores

3.5.8.3.1 - Árvores de regressão e classificação (CART)

Breiman, Freidman, Olshen e Stone desenvolveram o modelo CART em 1984(101). É um modelo que combina dois métodos baseados em árvores: regressão e classificação. O primeiro é um tipo de regressão não paramétrica para prever resultados contínuos com variáveis contínuas e/ou categóricas nas quais as informações são divididas em respostas binárias condicionais a perguntas que envolvem a variável preditiva. Já a árvore de classificação é muito semelhante a uma árvore de regressão, exceto pelo fato de ser uma classificação usada para prever uma resposta qualitativa em vez de quantitativa. Os modelos CART usam uma árvore binária para dividir repetidamente a variável preditiva em subconjuntos onde a distribuição da variável dependente é sucessivamente mais homogênea(102) (Figura 3).

Figura 3 – Diagrama representando a árvore CART.

A construção do conjunto de respostas binárias começa com a dicotomização de todos os preditores, exceto aqueles que já são originalmente binários. Assim, há uma categorização dicotômica de variáveis contínuas, e os preditores categóricos com mais de duas categorias são reorganizados para se tornarem binários. O processo de dicotomização de cada variável realizada pelo programa estatístico não é aleatório, mas baseado nos melhores pontos de corte, calculados pelo modelo, para melhor prever o risco alto ou baixo do resultado(102).

O segundo passo é identificar qual variável subsequente discrimina melhor a população com alto e baixo risco de desfecho, usando como parâmetros: significância, medidas de associação e área sob a curva (ASC) de cada preditor(102). O processo continua crescendo a árvore através das divisões até que o método encontre o melhor ponto em relação ao desempenho (Figura 4). Após esse ponto, com novas estratificações, as medidas de desempenho (significância, ASC, medidas de associação) serão melhores na amostra de treinamento, por excesso de adaptação (sobretraining), porém piores nas amostras de testes ou validação.

Figura 4 – Crescimento da árvore.

Em 2018, Manning *et al.* mostraram que é possível identificar pacientes com HNSCC com alto risco de readmissão no momento da alta inicial usando o CART. A presença de fistula foi o melhor parâmetro para distinguir se o paciente seria readmitido ou não. Entre os pacientes com fistula, a segunda melhor variável foi a renda bruta média ajustada pelo Código de Endereçamento Postal (CEP) e, entre os que não tiveram, o próximo melhor parâmetro foi o local primário do câncer(103) (Figura 5).

Figura 5 – Modelo CART para predição de readmissão hospitalar.

CART: *Classification and regression trees* (Árvores de regressão e classificação).

SBA: salário bruto ajustado; CEP: Código de Endereçamento Postal.

Adaptado de Manning *et al*(103).

3.5.8.3.2 - Floresta Aleatória

O método da floresta aleatória (“*random forest*”) é conhecido como um dos métodos de agrupamento, onde são construídas várias árvores, justificando o termo “floresta”(101). Cada árvore é obtida usando um determinado algoritmo e todas são diferentes uma da outra por dois motivos: primeiro porque em cada nó, a melhor divisão é escolhida a partir de um subconjunto aleatório de preditores e segundo porque cada árvore é construída usando uma parte da amostra (“*bootstrap sample*”). Uma parte dos dados (chamados em inglês de “*out-of-bag data*”) é usada apenas para estimar a precisão da previsão. Ao contrário de outros algoritmos baseados em árvores, não há poda de

árvores totalmente crescidas, para maximizar alguma medida de performance. Cada observação é atribuída a uma folha, o nó terminal de uma árvore, de acordo com a ordem e os valores das variáveis preditoras. Para uma árvore específica, as previsões para observações são fornecidas apenas para dados prontos para uso. A previsão geral é então calculada pela média de todas as árvores construídas.

3.5.8.3.3 - *Boosted regression trees* (BRT)

O BRT é uma técnica de regressão não paramétrica que combina uma árvore de regressão com um algoritmo de impulso. Essa variação da árvore de regressão clássica permite maior flexibilidade e desempenho preditivo na modelagem dos dados. Uma árvore de regressão particiona os dados de maneira binária, levando à criação de subamostras onde a distribuição da variável dependente é sucessivamente mais homogênea. Essas divisões são definidas por regras, métricas de distância ou ganho de informações. A seleção de variáveis e o valor no qual o ponto de divisão ocorre são determinados recursivamente em cada estágio da construção da árvore. O *boosting* melhora o aprendizado básico simples, ponderando novamente as observações que foram classificadas incorretamente ou que apresentaram erros residuais significativos na iteração anterior. Ou seja, cada fase melhora o desempenho do que não foi satisfatório na fase anterior. Quanto mais profundo crescemos a árvore, mais segmentos podemos acomodar e, portanto, mais variações podem ser explicadas. Isso resulta em maior complexidade do modelo e, portanto, um risco maior de ajustar o modelo aos dados. A motivação por trás do *boosting* é que, embora cada árvore tenha um poder preditivo superficial, o conjunto combinado desses classificadores fracos pode criar uma ferramenta de precisão robusta.

3.5.8.4 - Modelos Dinâmicos

Ao criar um modelo preditivo, é importante descrever detalhadamente em qual cenário ele foi concebido, ou seja, qual foi o momento utilizado para obter os dados preditivos, por exemplo: primeira consulta, pré-operatório, pós-operatório imediato etc. Prognósticos em câncer são frequentemente baseados em características conhecidas no

momento do diagnóstico e apresentadas como se fossem um veredicto vitalício. No entanto, quando um paciente com câncer sobrevive ao primeiro ano, seu prognóstico é possivelmente melhor do que no momento do diagnóstico. Isso é causado pelo fato de que eles sobreviveram ao primeiro período, onde em muitas análises preditivas parte dos pacientes morre.

Baseado na premissa supracitada, em 2012 foi publicado um estudo sobre um modelo prognóstico dinâmico em cabeça e pescoço. Os resultados mostraram de maneira significativa que os fatores prognósticos ocorridos durante o período de acompanhamento e após o tratamento inicial não são geralmente contemplados em modelos preditivos. Entretanto, foi demonstrado que incorporar tais mudanças, como sobreviver ao tratamento no primeiro ano, persistência tumoral, ou aparecimento de metástases adiciona um efeito importante no prognóstico do paciente portador de CECP(104). Por esse motivo, ao desenvolver uma nova ferramenta prognóstica, deve-se estar ciente de que tais previsões são limitadas ao tempo e ao cenário em que o paciente está envolvido. Variáveis que podem ser significativas em um momento para um determinado desfecho em um certo período de tempo, podem ser insignificantes em outra ocasião clínica.

3.5.9 - Validação de métodos

Julgar a acurácia de um modelo prognóstico apenas pelo conjunto de dados utilizados para a elaboração é um erro frequente. O teste mais rigoroso para qualquer modelo preditivo é uma validação externa, ou seja, a aplicação do modelo estabelecido em uma nova população. Muitas vezes, a falha de um modelo na validação externa poderia ter sido predita a partir de uma validação interna sem vieses. Em outras palavras, é provável que muitos modelos clínicos que não conseguiram validar em amostras externas, já teriam falhado em amostras parciais internas. Os principais métodos para obter validações internas com mais precisão são a divisão de dados, validações cruzadas e *bootstrapping*(66).

A divisão de dados consiste em dividir a amostra total em partes, reservando uma delas para uso exclusivo de validação. A grande vantagem é que esta amostra nunca é utilizada no desenvolvimento do modelo. A desvantagem é que o modelo será desenvolvido em uma amostra menor de pacientes, com menor poder. Este método pode

inviabilizar o desenvolvimento de modelos em amostras pequenas.

A validação cruzada é uma ferramenta de reamostragem usada para avaliar modelos preditivos e para validá-los internamente. O método possui um parâmetro único (k) que significa o número de grupos nos quais uma determinada amostra de dados deve ser dividida. Utiliza-se uma subamostra para estimar o desempenho do modelo quando aplicado em dados não utilizados durante o treinamento do modelo (105). As etapas da validação cruzada estão representadas na Figura 6 e descritas nos passos abaixo:

- 1) Reorganizar o conjunto de dados aleatoriamente.
- 2) Dividir o conjunto de dados em k subgrupos
- 3) Para cada rodada:
 - a. Escolher um subgrupo como conjunto de dados de validação
 - b. Utilizar os grupos restantes como um conjunto de dados de treinamento
 - c. Ajustar um modelo nos conjuntos de treinamento e avaliar no conjunto de validação
 - d. Salvar a pontuação da avaliação e descartar o modelo
- 4) Analisar a habilidade do modelo final usando o conjunto das pontuações de subgrupos

	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10
1ª										
2ª										
3ª										
4ª										
5ª										
6ª										
7ª										
8ª										
9ª										
10ª										

Figura 6 – Exemplo de esquema de validação cruzada em 10 subgrupos ($k=10$). Particionamento do conjunto de dados em subconjuntos mutuamente exclusivos, e posteriormente, o uso de alguns destes subconjuntos para a estimação dos parâmetros do modelo (dados de treinamento – em branco), enquanto os subconjuntos selecionados (em preto) são empregados na validação do modelo

Cada observação na amostra de dados é atribuída a um grupo individual e permanece nesse grupo durante o procedimento. Isso significa que cada amostra tem a oportunidade de ser usada no conjunto de dados de teste um vez e usada para treinar o modelo $k-1$ vezes(105).

O *bootstrapping* também é um método de reamostragem, que estima o erro padrão de sucessivas amostras isoladas (geralmente mais de 100 amostras de *bootstrapping*) e devolvidas ao conjunto de dados. O *bootstrapping* gera estimativas quase imparciais da precisão preditiva, que são de variação relativamente baixa e exigem menos ajustes do modelo do que a validação cruzada(66) (Figura 7).

Figura 7 – Representação gráfica do método de reamostragem *Bootstrapping* com o objetivo de obter uma distribuição normal de subamostras. Cada subamostra é devolvida ao conjunto de dados antes de nova seleção de subamostras.

Validações externas com populações distintas geográfica e temporalmente são ainda mais desejáveis, por inferir uma aplicabilidade quase universal de modelos preditivos. Datema *et al.* publicaram um estudo no qual um modelo preditivo de CECP criado na Holanda foi submetido a validação externa em pacientes americanos. O modelo mostrou resultados aceitáveis de calibração e discriminação na validação interna e externa, para a previsão de taxas de sobrevida em 2 e 5 anos (106).

3.5.10 - Ajuste

Em modelos preditivos deve-se evitar o ajuste excessivo, no qual o modelo se apresenta muito bem adaptado ao conjunto de dados que o compõe, porém se mostra ineficiente para prever novos desfechos em populações distintas, já que dificilmente teremos uma amostra perfeita de qualquer outra população(107) (Figura 8).

A complexidade e diversidade dos eventos biológicos pode gerar modelos preditivos tão detalhados e consequentemente específicos para uma determinada amostra, quanto mais completos forem. Além disso, é comum que a amostra disponível apresente desvios causados por fatores ao acaso ou por erros de medição. No momento em que o modelo se adapta a todos estes desvios, ocorre o sobreajuste. Para diminuir a possibilidade de ocorrer o sobreajuste deve-se penalizar o modelo pela sua complexidade,

com isto são eliminados parâmetros de pouca relevância, reduzindo a complexidade potencialmente deletéria. Uma maneira de se testar se está ocorrendo o sobreajuste é a utilização dos métodos de validação citados no item 3.5.9 ou de mecanismos de seleção de variáveis que penalizam o modelo pela sua complexidade, como o Akaike *Information Criteria* (AIC) (107).

Figura 8 – Representação gráfica de um modelo sobreajustado. A linha tracejada representa um modelo sobreajustado, enquanto a linha contínua representa um modelo adequado.

3.5.11 - Testando a performance de um modelo preditivo

Existem várias maneiras de avaliar o desempenho de um modelo preditivo. Nos últimos anos, muitos estudos propondo modelos de previsão de risco foram publicados, e a maioria das publicações utilizou a ASC como uma medida de qualidade, apesar de todas as críticas ao seu valor para esse fim (108). A ASC mede a capacidade de discriminação do modelo de maneira ordenada, mas não reflete o risco individual real dos pacientes.

O desempenho pode ainda ser quantificado em termos de calibração, usando, por exemplo, o teste Hosmer-Lemeshow "*goodness-of-fit*"(109). Trata-se de um teste que compara o risco previsto e o desfecho real dos pacientes em todas as subclasses de risco, por exemplo, quartis ou decimais de risco(109). Um valor p significativo sugere que as previsões de risco são significativamente diferentes do valor real, ou seja, calibração ruim (indesejada). Vale ressaltar que os critérios para divisão em subgrupos de risco podem ser bastante arbitrários(110) e certos pontos de corte podem fazer o modelo parecer mais

calibrado que outros. Idealmente, os pontos de corte devem ser escolhidos com base em critérios de relevância clínica e não para melhorar as medidas de calibração.

A calibração é uma medida de performance complementar à discriminação. É essencial verificar se os pacientes que foram previstos como tendo o desfecho para o resultado apresentam riscos mais altos do que aqueles que não foram, o que pode ser quantificado com medidas como sensibilidade, especificidade e ASC (ou estatística c de concordância)(111). O problema é que a ASC não distingue dois modelos - um preciso e outro que constantemente erra na mesma direção. Por exemplo, se um modelo for preciso e outro superestimar o risco individual de todos os pacientes em 10%, as curvas ROC serão igualmente discriminatórias, mas o primeiro modelo será mais calibrado que o segundo.

Outras medidas, como o coeficiente de determinação, também chamado de "R ao quadrado" (R^2), são usadas quase universalmente na avaliação de análises de regressão. O R^2 é uma medida do ajuste de um modelo estatístico linear generalizado, como regressão linear ou logística, em relação aos valores observados. O R^2 varia entre 0 e 1, indicando quanto o modelo pode explicar os valores observados. Quanto maior o R^2 , mais explicativo é o modelo, melhor ele se encaixa na amostra (112).

MÉTODOS

4 - MÉTODOS

4.1 – Seleção dos Indivíduos

Com o apoio financeiro da FAPESP e do Instituto Ludwig, em 2001 foi criado um grupo multicêntrico e multidisciplinar chamado GENCAPO. Entrevistadores treinados provenientes das diversas instituições participantes, dentre elas Instituto Arnaldo Vieira de Carvalho, Hospital Heliópolis, Hospital das Clínicas da Faculdade de Medicina da Universidade de São Paulo, Universidade do Vale do Paraíba e Faculdade de Medicina da Universidade de São Paulo de Ribeirão Preto coletaram prospectivamente dados demográficos, clínicos, patológicos e de seguimento dos pacientes portadores de CECP atendidos nestes centros. Tais informações foram obtidas segundo protocolos adotados por estudos epidemiológicos internacionais e armazenados em um banco de dados padronizado. Tratava-se, originalmente, de um estudo com dois braços principais: um estudo caso-controle, com pacientes portadores de CECP e pacientes de base hospitalar sem câncer, a fim de identificar fatores associados ao desenvolvimento de CECP; e uma coorte de pacientes com CECP, majoritariamente virgens de tratamento para a neoplasia.

Este estudo é uma análise retrospectiva de pacientes prospectivamente inseridos pelo grupo de pesquisa GENCAPO (<http://ctc.fmrp.usp.br/ClinicalGenomics/cp/>) na coorte de pacientes com CECP. Foram analisados pacientes incluídos no banco de dados entre 04 de julho de 2000 e 16 de agosto de 2011, com diagnóstico de carcinoma epidermoide de cavidade oral, orofaringe, hipofaringe ou laringe. Não foram incluídos os pacientes portadores de carcinoma de cavidade nasal, nasofaringe, seios paranasais, lábios e tumores primários ocultos.

4.2 – Ética

Esse estudo é um subprojeto do Projeto “FATORES AMBIENTAIS, CLÍNICOS, HISTOPATOLÓGICOS E MOLECULARES ASSOCIADOS AO DESENVOLVIMENTO E AO PROGNÓSTICO DE CARCINOMAS EPIDERMÓIDES DE CABEÇA E PESCOÇO” desenvolvido no departamento de Cirurgia do Hospital das

Clínicas da FMUSP e previamente aprovado pela CAPPesq sob o nº 0511/07 (Apêndice 1).

Utilizamos os dados dos pacientes do subprojeto “BUSCA DE MARCADORES DE AGRESSIVIDADE EM TUMORES DE CABEÇA E PESCOÇO” desenvolvido no departamento de Cirurgia do Hospital das Clínicas da FMUSP e previamente aprovado pela CAPPesq sob o nº 0511/07. (Apêndice 2).

Todos os pacientes incluídos neste estudo assinaram o Termo de Consentimento livre esclarecido no momento de sua inclusão no projeto.

4.3 – Preenchimento dos questionários Gencapo

Uma vez integrante do Gencapo, todos os pacientes foram submetidos a questionários aplicados por entrevistadores treinados os quais preenchem formulários padronizados. Informações pessoais, profissionais, de familiares e hábitos relacionados ao tabagismo e etilismo dos indivíduos eram registradas (Apêndice 3). Os formulários clínicos foram preenchidos pelos médicos cirurgiões responsáveis pelo atendimento dos pacientes (Apêndice 4). As informações provenientes dos laudos histopatológicos foram coletadas por médicos pertencentes ao Gencapo (Apêndice 5). Os questionários de acompanhamento foram preenchidos durante o retorno ambulatorial ou hospitalar do paciente, bem como em qualquer outro momento caso houvesse informação de óbito (Apêndice 6). Todas as informações constantes nos formulários físicos foram posteriormente incluídas na plataforma informatizada online do GENCAPO.

4.4 – Seleção das Variáveis

Os dados disponíveis foram obtidos no momento da apresentação e diagnóstico do paciente: sexo, idade, raça, escolaridade, tabagismo, etilismo, presença de sintomas (alterações visuais, dificuldade respiratória, traqueostomia, disfagia, disfonia, dor, feridas, lesões, nódulos, obstrução nasal, odinofagia, otalgia, rinorréia, sangramento, sensação de corpo estranho, tosse), índice de massa corpórea (IMC), porcentagem da perda ponderal em relação ao peso habitual, tempo de perda ponderal, fração dos

familiares de primeiro grau que tiveram câncer, localização da lesão (através do Código Internacional de Doenças), extensão da doença aferida por avaliação clínica e/ou radiológica (tamanho da lesão primária, extensão linfonodal da doença e metástase à distância). O grau de diferenciação tumoral, apesar de poder ser aferido com aceitável acurácia em materiais de biópsias, foi obtido por relatórios patológicos tanto de biópsias como de peças cirúrgicas, conforme disponibilidade. O acompanhamento durante cinco anos com o desfecho (cura ou óbito, independentemente de sua causa) foi realizado conforme retornos ambulatoriais dos pacientes, bem como por busca ativa do time de pesquisadores, por diversos métodos, dentre eles bases oficiais de registros de óbito, nos casos de perda de seguimento.

Todos os sintomas supracitados foram submetidos a uma regressão logística passo a passo (*stepwise regression*) prevendo mortalidade como desfecho. Após treinamento e validação, selecionamos quatro sintomas: nódulos, disfonia, lesões e odinofagia. Nódulos já são contemplados pelo TNM assim como a presença de lesões, portanto ambas foram excluídas do modelo, restando apenas disfonia e odinofagia como sintomas relevantes para inclusão nos modelos de predição posteriormente desenvolvidos.

Após realizar a primeira triagem das variáveis candidatas, analisamos a quantidade e distribuição dos dados faltantes entre elas (Tabela 3).

Tabela 3 – Número e proporção dos dados faltantes por variáveis.

Variável	N	Dados faltantes	Dados faltantes (%)
Antecedente familiar de cancer	1793	299	14,3
cT	2009	83	4,0
cN	1979	113	5,4
cM	1762	330	15,8
Maior Dimensão Clínica (DC1)	1473	619	29,6
Menor Dimensão Clínica (DC2)	1335	757	36,2
Espessura Clínica	857	1235	59,0
Disfonia	2027	65	3,1
Escolaridade	2042	50	2,4
Etilismo	2026	66	3,2
Grau de Diferenciação	1925	167	8,0
Idade	2092	0	0,0
IMC	1816	276	13,2
Odinofagia	2014	78	3,7
Perda Ponderal (%)	1904	188	9,0
Perda Ponderal (Kg)	1967	125	6,0
Perda Ponderal (m)	1176	1016	48,7
Perda Ponderal (S/N)	2005	87	4,2
Raça	2063	29	1,4
Sexo	2092	0	0,0
Sítio	2092	0	0,0
Status 5 anos (vivo/morto)	1811	281	13,4
Tabagismo	2035	57	2,7
Volume Tumoral	855	1237	59,1

cT: estágio clínico do tumor primário; cN: estágio clínico dos linfonodos regionais; cM: estágio clínico de metástase à distância; m: meses; S/N: sim ou não.

4.5 – Material Biológico

Os blocos parafinados produto das ressecções cirúrgicas ou biópsias foram obtidos após solicitação junto ao Serviço de Anatomia Patológica da FMUSP. Os blocos foram analisados para seleção daqueles com maior representação de tecido e armazenados no Laboratório de Investigação Médica (LIM)28.

4.6 – Pesquisa do Gene do Papilomavírus Humano

A pesquisa de ácido desoxirribonucleico (DNA) do HPV por reação em cadeia da polimerase (PCR) foi escolhida como método de detecção, pois utilizamos amostras embebidas em parafina, mais facilmente disponíveis.

4.7 – Extração de DNA de tecido parafinado

Uma vez de posse dos blocos de parafina, três cortes com dez micrômetros cada foram obtidos através do uso do micrótomo histológico. Cuidados padrões, como a antisepsia entre cada bloco (álcool 70%), foram tomados para evitar a contaminação cruzada pelo HPV entre amostras

O estudo de HPV foi realizado em duas fases:

A primeira, com material de 350 pacientes com tumores de diversos sítios, foi realizada anteriormente ao presente estudo pelo time do Dr José Eduardo Levi pelo GENCAPO, em parceria com consórcio internacional. Utilizou-se o método INNO-LiPA (Innogenetics) cuja tecnologia é baseada na hibridização reversa capaz de detectar simultaneamente 28 tipos de HPV, conforme previamente publicado (16, 113).

A segunda, foi efetivada especificamente para o presente estudo, na Fundação Oncocentro de São Paulo (FOSP) em 342 pacientes utilizando o teste BD Onclarity HPV Assay[®]. Trata-se de um ensaio em tempo real no qual é possível detectar informações genéticas de seis genótipos do HPV (16, 18, 31, 45, 51 e 52), além de relatar outros oito genótipos em três grupos (33/58, 56/59/66 e 35/39/68). É o primeiro ensaio que detecta especificadamente os genes responsáveis para oncoproteínas E6 e E7 (114, 115), sendo realizado de forma 100% automatizada através do sistema BD Viper LT[®]. Este ensaio não prevê o teste de material fixado com formol e embebido em parafina, portanto para a genotipagem destas amostras, foi preciso cumprir algumas etapas pré-analíticas, manualmente: (1) realização de 3 cortes histológicos, com espessura de 10µm, colocados em um microtubo; (2) identificação com etiquetas dos tubos BD Onclarity[®] HPV LBC em que foram colocadas as amostras; (3) passagem dos 3 cortes do microtubo para o tubo BD Onclarity[®] HPV LBC; (4) adição de 0,5 ml de água destilada; (5) inversão do tubo com o material parafinado dentro 3 a 4 vezes para garantir a mistura do material com o diluente. A amostra foi então submetida a lise celular com consequente exposição do DNA a ser extraído e amplificado, utilizando a estação preaquecedora Viper LT[®]. Em seguida, foi transferida para uma outra parte do aparelho no qual foi submetida ao processamento automático visando análise através do PCR. O DNA foi extraído usando partículas magnéticas BD FOX[®] e o eluente contendo DNA foi usado para pesquisar três

reações de genotipagem de PCR: G1 detecta HPV 16, HPV 18 e HPV 45; G2 detecta HPV 31, HPV 33 $\pm$ 58 e HPV 56-59-66, o G3 detecta HPV 51, HPV 52 e HPV 35-39-68. Em todas reações foram analisadas a presença do gene endógeno da betaglobina humana buscando controlar a presença de inibidores da PCR, avaliação da adequação, extração e amplificação da amostra. Após 40 ciclos de PCR, consideramos positiva qualquer pontuação do escore do liminar do ciclo (Ct) para um tipo específico e/ou controle interno da betaglobina reagente(116).

Esse procedimento foi realizado com 342 amostras. O relatório do Viper apresentou falha no controle interno (ICt) em 43 amostras. Os dados brutos destas amostras tidas como ICt foram extraídas do sistema BD Viper LT[®] através de um pen drive, e enviadas para a BD dos Estados Unidos da América para serem analisadas. Essa análise mostrou que apesar do resultado IC, as amostras eram positivas para beta globina (gene humano) e em uma delas havia inclusive material genético de HPV 16. Isso aconteceu pelo fato de o equipamento ter sido parametrizado para uso com citologia líquida, e não para material fixado em formol e embebido em parafina. Dessa maneira, o DNA existente na amostra não alcançava o limiar de fluorescência pré-estabelecido, após os ciclos de PCR.

Classificamos como positivo para pesquisa do HPV todas as amostras com subtipos considerados de alto risco (HPV16, 18, 31, 33, 35, 39, 45, 51, 52, 56, 58, 59, 66, 68, 73)(117, 118).

4.8 – Métodos estatísticos

4.8.1 – Imputação de dados

4.8.1.1 Dimensões Tumorais no estudo de anatomia patológica.

Apesar de não incluídos dados de relatórios de anatomia patológica (AP) em nossa análise, já que pretendemos criar um modelo para aplicação previamente ao tratamento, tais dados foram utilizados para elaboração de métodos de imputação das informações faltantes de dimensões e espessura clínicas. Os dados faltantes das dimensões AP foram imputados utilizando a média por sítio em cada uma das classes de estágio pT. Havia três casos T₁ de Hipofaringe que não possuíam informação da terceira dimensão patológica,

portanto não foi possível calcular a média. Para imputar estes três dados utilizamos a média de duas regressões lineares bivariadas (dimensão AP1 e dimensão AP2).

4.8.1.2 - Dimensões tumorais clínicas

Os dados das dimensões tumorais clínicas sofreram imputação, sendo inicialmente realizada por regressão linear em função das dimensões registradas do tamanho tumoral no estudo de anatomia patológica para os pacientes que tinham tais informações. Quando ausentes, utilizamos a média dos dados disponíveis em cada categoria de estágio cT e por cada sítio tumoral. Por fim, nos casos em que nenhum dos métodos conseguiu imputar as dimensões clínicas, utilizamos a técnica de contração (*Shrinkage estimator*) para covariâncias.

Após a imputação das três dimensões clínicas tumorais, criamos uma variável, chamada de volume tumoral. Para este cálculo, utilizamos a fórmula descrita por Monga $[(\text{comprimento (DC1)} \times \text{largura (DC2)} \times \text{profundidade (espessura)})/2)](119)$.

4.8.1.3 – Sítios dos tumores primários

Após a análise inicial dos dados, em 31 pacientes havia dúvidas se o tumor havia se originado na orofaringe ou na cavidade oral (Figura 9).

Figura 9 – Distribuição pelos sítios antes da realocação do Cavidade Oral/Orofaringe.

Para definir o sítio destes 31 indivíduos realizamos um modelo CART apenas considerando o sítio e o desfecho (Figura 10), a fim de prever se os casos indefinidos se comportavam predominantemente como tumores de cavidade oral ou de orofaringe.

Figura 10 – CART analisando o sítio e o desfecho, a fim de melhor alocar os casos sem definição de sítio entre cavidade oral e orofaringe

Analisando a árvore da figura 10, observamos o agrupamento da categoria “cavidade oral/orofaringe” com “orofaringe”. Com todas as variáveis de interesse, o sítio não foi selecionado no início da árvore, porém, quando foi, apareceu ao final da árvore, além de agrupar novamente os casos de indefinição juntamente à orofaringe e não à cavidade oral. Deste modo, classificamos os 31 casos sem distinção de sítio entre cavidade oral/orofaringe como sendo casos de câncer de orofaringe (OF) (Figura 11).

Figura 11 - Distribuição pelos sítios após a realocação

4.8.1.4 – Perda Ponderal (meses)

Analisando a variável “Perda Ponderal (m)” pela variável binária (Perda Ponderal (S/N)), constatamos que em 100 pacientes houve perda ponderal, mas os dados de Perda Ponderal (m) estavam faltando.

Decidimos pela imputação através de uma regressão linear bivariada utilizando como variável independente a “Perda ponderal (kg)” (Figura 12)

Figura 12 – Regressão linear bivariada.

Fórmula utilizada na imputação dos pacientes que são SIM para Perda Ponderal (S/N), mas estavam com dados faltantes para Perda Ponderal (m)

$$(1) \text{ Perda Ponderal (m)} = 4,2311457 + 0,1474558 * \text{Perda Ponderal (Kg)}$$

A distribuição pós imputação dos dados de Perda Ponderal (m) está representada na Figura 13.

Figura 13 - Distribuição da Perda Ponderal (m).

4.8.2 – Análise Estatística

Todas as informações contidas no banco de dados do Gencapo foram importadas com o auxílio do programa Excel (Microsoft® Office Professional 365). Diversos modelos de predição foram usados para investigar os efeitos dessas variáveis simultaneamente na sobrevida global e para desenvolver o modelo de risco de cada indivíduo. Estratégias de mineração de dados, como árvores de classificação e regressão, *boosted tree* e floresta aleatória foram usadas para explorar interações entre variáveis, a fim de tornar os modelos preditivos mais precisos. As performances dos diferentes modelos preditivos gerados foram analisadas através do R^2 , da área sob a curva ROC (ASC) para discriminação (120), e Hosmer-Lemeshow *goodness-of-fit statistic* para calibração (109).

Um modelo preditivo de redes neurais e outro baseado em regressão logística foram criados para todos os sítios e para cada um dos sítios. Subdividimos nossa população em quartis de acordo com a faixa de risco de mortalidade. Os grupos foram denominados: colocar as faixas de corte para cada um.

Estes modelos foram comparados aos modelos baseados no TNM 7ª e 8ª edição utilizando curvas de sobrevida de Kaplan Meyer.

Para avaliar o incremento do HPV em predição de mortalidade, desenvolvemos modelos de regressão logística que incluíam apenas pacientes com informação do status HPV, comparando-se o modelo gerado apenas com informações clínicas ao modelo gerado com informações clínicas e o HPV.

Para todas as análises estatísticas utilizamos o JMP® Pro 10.0.0 e o Excel (Microsoft® Office Professional 365).

4.8.3 – Validação

Para validação optamos pela estratégia de divisão da amostra de pacientes em três subconjuntos: treinamento, validação e teste. O primeiro é a parte usada para estimar os parâmetros do modelo. O segundo é a parte que avalia ou valida a capacidade preditiva do modelo. O terceiro é uma avaliação final independente da capacidade preditiva do modelo.

Criamos duas variáveis: “data3” e “data3.50.25.25”. No caso da primeira, geramos três grupos de forma aleatória com número semelhante de pacientes. Já a segunda foi realizada também de forma aleatória, mas a quantidade de pacientes foi de 50% para o grupo treinamento, 25% para validação e 25% para teste. Para geração das sequências aleatórias com as duas diferentes proporções, utilizamos o software Excel (Microsoft® Office Professional 365).

4.8.4 – Transformação de variáveis contínuas

Com o objetivo de testar inúmeras variáveis durante a construção dos modelos preditivos, foi realizado a categorização das seguintes variáveis contínuas: idade, perda ponderal (kg), perda ponderal (m) e volume clínico. Além disso testamos a transformação logarítmica da variável “volume clínico”, para minimizar o efeito deletério no modelo de regressão logística dos valores extremos observados para nesta variável.

4.8.4.1 – Idade

Utilizamos o CART para definir o valor de corte para categorizar de forma binária a variável idade. Conforme a figura 14, o valor encontrado foi 76 anos. Para fins práticos,

resolvemos adotar as seguintes categorias: menor que 75 anos e maior ou igual a 75.

Figura 14 – CART utilizado para a criação de categorias para a variável “Idade”.

A distribuição da nova variável para idade encontra-se na figura 15.

Figura 15 – Distribuição para idade categórica.
(não se <75 anos; sim se ≥ 75 anos).

4.8.4.2 – Perda Ponderal

Duas variáveis relacionadas à perda ponderal foram dicotomizadas: perda ponderal em quilos e tempo da perda ponderal em meses. Da mesma maneira que fizemos com “Idade”, também utilizamos CART para definir os valores de corte.

4.8.4.3 – Volume Clínico

Mais uma vez utilizamos o CART para definir os valores de corte, porém, diferentemente das anteriores, criamos duas variáveis categóricas de volume clínico,

sendo uma com dois níveis e outra com quatro níveis (Tabela 4).

Tabela 4 – Categorização da variável “volume clínico” (valores em cm^3).

	Categorias			
	1	2	3	4
Volume Clínico 2 categorias	≤ 25	> 25		
Volume Clínico 4 categorias	$\leq 1,0$	$> 1,0$ $\leq 5,0$	$> 5,0$ $\leq 25,0$	> 25

A distribuição do volume clínico para tumores laríngeos apontava para um ponto de corte menor que os demais sítios. Após análise utilizando um modelo CART criamos uma variável categórica binária exclusiva para laringe: categoria 1 para tumores menores ou iguais a 3cm^3 e categoria 2 para tumores maiores que 3cm^3 .

Além disso, devido a presença de valores extremos, resolvemos criar a variável logarítmica do “volume clínico” para todos os sítios (Tabela 5).

Tabela 5 – Comparação da distribuição entre “volume clínico” e “volume clínico LOG”.

Volume Clínico			Volume Clínico LOG

	Quartis			Quartis	
100%	máximo	472,5	100,0%	máximo	2,67
75%	quartil	22,64	75,0%	quartil	1,36
50%	mediana	8,66	50,0%	mediana	0,94
25%	quartil	3	25,0%	quartil	0,48
0%	mínimo	0	0,0%	mínimo	-2,40
Média		20,80	Média		0,88

LOG: logaritmo

4.8.5 – Criação da variável linfonodomegalia

Utilizando informações de cN, foi criada a variável binária “linfonodomegalia” (não se cN0; sim para os demais valores de cN)

4.9 – Variáveis utilizadas

Segue abaixo a lista final com todas as 27 variáveis que foram utilizadas para construção dos modelos preditivos:

- Dimensão Clínica 1
- Dimensão Clínica 2
- Disfonia
- Escolaridade
- Espessura clínica
- Etilismo
- Grau de diferenciação
- Idade
- Idade Categórica (sim se ≥ 75 anos)
- IMC
- Linfonodomegalia
- Metástases (distância)
- Odinofagia
- Perda ponderal (%)
- Perda ponderal (kg)
- Perda ponderal (kg) Categórica (sim se $\geq 3,5$ kg)
- Perda ponderal (m)
- Perda ponderal (m) categórica (sim se ≥ 5 meses)
- Perda ponderal (sim/não)
- Raça
- Sexo
- Sítio
- Tabagismo
- Volume Clínico
- Volume Clínico Ordinal (cat1 $\leq$ 1,0) (1,0<cat2 $\leq$ 5,0) (5,0<cat3 $\leq$ 25,0) (cat4>25)
- Volume Clínico Binário (cat1 $\leq$ 25,0) (cat2>25)
- Volume Clínico LOG
- Volume Clínico Binário LA (cat1 $\leq$ 3,0) (cat2>3,0) – específica para tumores laríngeos

RESULTADOS

5.0 – RESULTADOS

Os principais achados estão dispostos em etapas conforme esquema a seguir, enquanto os achados adicionais constam nos anexos indicados.

- Descrição da amostra de pacientes incluídos
- Modelos preditivos com todos os subsítios anatômicos
 - Comparação com o sistema TNM
- Modelos preditivos em cavidade oral
 - Comparação com o sistema TNM
- Modelos preditivos em orofaringe
 - Comparação com o sistema TNM
- Modelos preditivos em hipofaringe
 - Comparação com o sistema TNM
- Modelos preditivos em laringe
 - Comparação com o sistema TNM
- Incremento da performance dos modelos preditivos após adição do *status* HPV

No período entre 04 de julho de 2000 e 16 de agosto de 2011, o banco de dados do GENCAPO recebeu dados de 2930 pacientes, dos quais 838 pacientes foram excluídos por um ou mais dos seguintes motivos: ausência de informação do tumor primário, ausência de confirmação histológica do diagnóstico de carcinoma epidermóide, tratamento oncológico prévio e/ou tumores com variantes verrucosos. Restaram 2092 pacientes, sendo que 281 pacientes não possuíam informações sobre seguimento de pelo menos cinco anos após a entrevista. Estes pacientes foram excluídos das análises estatísticas. Desta forma, 1811 pacientes estavam elegíveis para o desenvolvimento dos modelos preditivos.

5.1 – Descrição da amostra total de pacientes

Os dados sociodemográficos, clínicos e de acompanhamento dos 2092 pacientes inseridos encontram-se no Anexo A, Tabelas A.1 a A.7. O perfil dos pacientes incluídos neste estudo é formado em sua maioria por homens brancos, em média na sexta década de vida, predominantemente tabagistas e com escolaridade média até o quinto ano do ensino fundamental.

Ao analisar a distribuição de acordo com o sítio, observamos que pouco mais de quarenta por cento da nossa população é composta por tumores cujo sítio primário é a cavidade oral, seguidos respectivamente por tumores da laringe, orofaringe e hipofaringe (Tabela 6).

Dados relacionados ao tamanho tumoral evidenciam uma população com tumores avançados, sendo que mais de um terço dos pacientes foram classificados como estágio T avançado pelo sistema TNM desde a primeira consulta (Tabela 6).

Já em relação ao grau de diferenciação, a maioria dos pacientes tiveram o diagnóstico de carcinoma epidermoide moderadamente diferenciado e quase dez por cento foram classificados como pouco diferenciados (Tabela 6).

Analisando a tabela 6 é possível inferir que os dados faltantes não diferem de forma clinicamente relevante da população de estudo, no que se refere à distribuição por sítio tumoral, estadiamento TNM ou grau de diferenciação.

Tabela 6 – Distribuição de acordo com o sítio, cT, cN, M e o grau de diferenciação tumoral dos pacientes incluídos e excluídos das análises.

Característica		População GENCAPO n (%)	Desfecho faltante n (%)	Incluído nas análises n (%)
Sítio tumoral	Cavidade Oral	843 (40,3)	91 (32,3)	752 (41,5)
	Orofaringe	444 (21,2)	46 (16,4)	398 (22,0)
	Hipofaringe	199 (9,5)	32 (11,4)	167 (9,2)
	Laringe	606 (29,0)	112 (39,9)	494 (27,2)
T	1	252 (12,6)	49 (19,5)	203 (11,6)
	2	479 (23,9)	77 (30,7)	402 (23,0)
	3	445 (22,2)	64 (25,5)	381 (21,8)
	4a	798 (39,9)	61 (24,3)	737 (42,1)
	4b	27 (1,4)	0 (0)	27 (1,5)
N	0	901 (45,5)	144 (59,3)	757 (43,6)
	1	286 (14,4)	38 (15,6)	248 (14,3)
	2a	121 (6,1)	16 (6,6)	105 (6,1)
	2b	278 (14,0)	18 (7,4)	260 (15,0)
	2c	251 (12,7)	16 (6,6)	235 (13,5)
	3	142 (7,1)	11 (4,5)	131 (7,5)
M				1795
	0	2074 (99,1)	279 (99,3)	(99,1)
	1	18 (0,9)	2 (0,7)	16 (0,9)
Grau de Diferenciação	Bem diferenciado	641 (32,3)	75 (29,5)	566 (33,9)
	Moderadamente	1104 (57,4)	161 (63,3)	943 (56,4)
	Pouco diferenciado	180 (9,4)	18 (7,1)	162 (9,7)

A distribuição dos tamanhos tumorais em três dimensões e do volume tumoral encontram-se na tabela 7 e 8 respectivamente. Nota-se que a amostra é composta de pacientes com tumores volumosos, em média com 4cm na maior dimensão e 20cm³.

Tabela 7 – Distribuição dos tamanhos tumorais em três dimensões.

		Dimensão 1	Dimensão 2	Espessura
100.0%	máximo	15	10,0	8,0
99.5%		12	9,0	6,0
97.5%		10	6,4	5,0
90.0%		6,5	5,0	3,1
75.0%	quartil	5	3,9	2,6
50.0%	mediana	4	2,8	2,0
25.0%	quartil	2,7	2,0	1,1
10.0%		2	1,0	0,8
2.5%		1	0,7	0,5
0.5%		0,5	0,4	0,2
0.0%	mínimo	0,2	0,1	0
Média		4,11	2,86	2,01
Desvio Padrão		2,06	1,50	1,14
Erro de Desvio Padrão		0,05	0,03	0,03
Superior 95% da média		4,20	2,93	2,06
Inferior 95% da média		4,01	2,79	1,95
N		1809	1809	1809

Tabela 8 – Distribuição do volume tumoral (em cm³)

100.0%	máximo	472,5
99.5%		298
97.5%		124,4
90.0%		45
75.0%	quartil	22,6
50.0%	mediana	8,7
25.0%	quartil	3
10.0%		1
2.5%		0,25
0.5%		0,04
0.0%	mínimo	0
Média		20,80
Desvio Padrão		39,44
Erro de Desvio Padrão		0,93
Superior 95% da média		22,61
Inferior 95% da média		18,98
N		1809

Tivemos quatro variáveis relacionadas ao peso do paciente: perda ponderal em quilos, tempo da perda ponderal em meses, percentual da perda ponderal e IMC. Todos estes dados estão em tabela no Anexo A, Tabelas A.5, A.6 e A.7. Cabe ressaltar que 1073 (61%) pacientes cursaram com perda ponderal, sendo que em 843 pacientes esta perda foi superior a cinco quilos.

5.2 – Modelos preditivos para tumores de todos os sítios

5.2.1 – Modelos baseados em árvores

5.2.1.1 – Modelo CART para tumores de todos os sítios

Empregando todas as variáveis citadas no item 4.9, foi construído o modelo CART para todos os 1811 pacientes com seguimento completo. Para realizar validação do modelo, usamos o DATA3 (item 4.8.3). O melhor valor de R^2 foi definido como o parâmetro para término da construção da árvore.

Na tabela 9 e na figura 16 encontram-se informações a respeito dos testes e da sequência de divisões.

Tabela 9 – Medidas de R^2 e número de pacientes para cada subdivisão da amostra - modelo CART para todos os sítios tumorais.

	R^2	N	Número de divisões	Atribuições	AIC
Treinamento	0,105	630	4	23	1889,39
Validação	0,063	574			
Teste	0,054	607			

AIC: Akaike information criterion

R^2 : coeficiente de determinação

Figura 16 – Comportamento do R^2 ao longo da sequência de divisões nas diferentes subdivisões da amostra de pacientes.

Dados de treinamento em azul; dados de validação em vermelho; dados do teste em laranja. O conjunto de treinamento é a parte usada para estimar os parâmetros do modelo. O conjunto de validação é a parte que avalia ou valida a capacidade preditiva do modelo. O conjunto de testes é uma avaliação final independente da capacidade preditiva do modelo.

Analisando a figura 17, observamos a sequência de divisões que sugere uma possível interação entre volume clínico e linfonodomegalia.

Figura 17 – Árvore de classificação com todos os sítios, modelo CART

A tabela 10 demonstra a discriminação do modelo desenvolvido pelo método de CART, através das curvas ROC (e respectivas ASCs calculadas) das subdivisões em treinamento, validação e teste.

Tabela 10 – Curvas ROC do modelo CART para todos os sítios.

ASC: área sob a curva

Entre todas as 27 variáveis preditivas utilizadas na construção do modelo CART, apenas três contribuíram para o modelo final (Tabela 11).

Tabela 11 – Contribuição das variáveis no modelo CART para todos os sítios.

Variável	Número de divisões	G ²	G ²
Volume Clínico	1	48,9	<div style="width: 100%;"></div>
Linfonodomegalia	2	24,4	<div style="width: 50%;"></div>
Sexo	1	9,4	<div style="width: 25%;"></div>

AIC: Akaike information criterion

5.2.1.2 – *Boosted Regression Trees* para tumores de todos os sítios

Empregando todas as variáveis descritas no item 4.9, construímos o modelo pelo método de BRT para todos os sítios. Utilizamos 100 camadas com número mínimo de 3 divisões e a variável DATA3 para validação.

Dentre as 27 variáveis, 23 contribuíram na elaboração final do modelo baseado em BRT. (Tabela 12). Sexo, linfonodomegalia e Idade Categórica (com ponto de corte em 75 anos) foram as variáveis que mais determinaram a construção deste modelo ($R^2=0,082$) e a ASC em uma subamostra para teste foi de 0,75 (Tabela 13).

Tabela 12 – Contribuição das variáveis preditivas para construção do modelo de BRT para todos os sítios.

Variável	Número de divisões	G ²	G ²
Sexo	26	16423,9	
Linfonodomegalia	18	13048,3	
Idade (sim se ≥75a)	9	7721,0	
Etilismo	11	6779,5	
Perda Ponderal (m) (sim se ≥5m)	9	4762,5	
Volume Clínico Ordinal	8	4438,7	
Perda Ponderal (Kg)	5	3739,0	
Idade	11	3320,6	
Volume Clínico	7	3247,7	
Sítio	5	3132,1	
Grau de Diferenciação	2	2261,2	
Espessura Clínica	4	2257,6	
Volume Clínico Binário	6	2107,1	
Odinofagia	11	1534,0	
Perda Ponderal (S/N)	18	1506,2	
Perda Ponderal (m)	2	921,6	
Dimensão Clínica 1	1	628,2	
Tabagismo	3	477,8	
IMC	1	466,3	
Disfonia	2	351,4	
Perda Ponderal (%)	1	130,6	
Dimensão Clínica 2	1	23,6	
Escolaridade	1	8,7	

a: anos; m: meses; cat: categoria em cm³; S/N: sim ou não

IMC: índice de massa corpórea

Volume clínico binário (cat1≤25 cm³) (cat2>25 cm³)

Volume clínico ordinal (cat1≤1,0 cm³) (1,0 cm³<cat2≤5,0 cm³) (5,0 cm³<cat3≤25,0 cm³) (cat4>25 cm³)

Tabela 13 – Curvas ROC do modelo de BRT para todos os sítios.

ASC: área sob a curva

5.2.1.3 – Floresta Aleatória para tumores de todos os sítios

Elaboramos um modelo de Floresta Aleatória utilizando como parâmetro a construção de 100 árvores com um mínimo de 10 divisões em cada árvore. Utilizamos o DATA3 como validação.

Dentre todas as 27 variáveis, a única que não contribuiu na construção de nenhuma árvore foi a presença de metástase à distância. A contribuição de todas as variáveis e discriminação do modelo encontram-se na Tabela 14 e 15. Neste modelo as três mais importantes foram linfonodomegalia, maior dimensão clínica e volume clínico logarítmico.

Tabela 14 - Contribuição das variáveis preditivas para construção do modelo de Floresta Aleatória para todos os sítios.

Variável	Número de divisões	G ²	G ²
Linfonodomegalia	137	1311,4	
Dimensão Clínica 1	136	1171,9	
Volume Clínico LOG	133	1048,2	
Volume Clínico	141	1040,7	
Dimensão Clínica 2	131	820,3	
Espessura Clínica	114	811,6	
Escolaridade	125	719,7	
Perda Ponderal (Kg)	106	697,2	
Idade	135	678,1	
Etilismo	144	608,4	
Volume Clínico Ordinal	105	608,3	
IMC	114	517,3	
Perda Ponderal (%)	93	468,7	
Grau de Diferenciação	106	389,2	
Perda Ponderal (m) (sim se ≥5m)	107	366,4	
Sítio	97	318,9	
Perda Ponderal (Kg) (sim se ≥3,5kg)	71	296,1	
Tabagismo	82	279,9	
Perda Ponderal (S/N)	71	258,3	
Raça	79	251,1	
Volume Clínico Binário	65	224,8	
Odinofagia	104	216,1	
Perda Ponderal (m)	67	205,6	
Disfonia	90	193,4	
Sexo	73	190,0	
Idade (sim se ≥75a)	34	80,9	

LOG: logaritmo; IMC: índice de massa corpórea; m: meses; a: anos; cat: categoria; Volume clínico ordinal (cat1≤1,0cm³) (1,0 cm³<cat2≤5,0 cm³) (5,0 cm³<cat3≤25,0 cm³) (cat4>25 cm³)
 Volume clínico binário (cat1≤25 cm³) (cat2>25 cm³)

Tabela 15 – Curvas ROC do modelo de Floresta Aleatória para todos os sítios.

ASC: área sob a curva

5.2.2 - Redes Neurais para tumores de todos os sítios

Todas as variáveis listadas no item 4.9 foram utilizadas para construção do modelo baseado em redes neurais. Para validação utilizamos o “data3.50.25.25”. Definimos três nós em uma única camada. No Anexo B encontra-se a Tabela B.1 com informações a respeito dos três grupos (treinamento, validação e teste).

Com relação à discriminação, observa-se que mesmo sendo testado em uma amostra que não foi utilizada na construção, a ASC foi de 0,76 (Tabela 16).

Tabela 16 – Curvas ROC do modelo de redes neurais para todos os sítios.

ASC: área sob a curva

5.2.3 – Regressão Logística para tumores de todos os sítios

Segundo Harrell *et al*, é preciso checar o número de desfechos menos frequentes antes de definir a quantidade de variáveis preditivas que entrarão no modelo. Utilizamos a relação de 10 eventos por variável. Na nossa amostra, 1163 pacientes faleceram antes de completar o quinto ano de acompanhamento e 648 sobreviveram. Portanto é possível a construção de um modelo utilizando no máximo 64 variáveis sem comprometer a estabilidade do mesmo.

Análises preliminares utilizando modelos baseados em árvores apresentaram possíveis fatores de valor preditivo e, associado ao conhecimento clínico, algumas interações possíveis. Todas as interações foram testadas, mas não apresentaram valores preditivos significativos. Vários modelos de regressão logística foram construídos até chegarmos a este modelo final, de melhor performance.

Realizamos a seleção dos preditores utilizando como método de parada o *Max Validation R^2* , com direção anterógrada combinada, sendo selecionadas pelo *software* como variáveis com coeficientes estatisticamente significativos a idade, linfonodomegalia, sexo e volume clínico ordinal (quatro categorias). As demais variáveis que entraram no modelo foram: etilismo, grau de diferenciação, idade categórica (sim se $\geq 75a$), IMC, odinofagia, perda ponderal (%), perda ponderal (Kg), perda ponderal (S/N) e sítio. (Tabela 17). Dos 1811 elegíveis, foram excluídos das análises 358 por não possuírem dados completos de todas as variáveis. 1453 pacientes foram divididos em três grupos para treinamento (n=722), validação (n=348) e teste (n=383) (Anexo B, Tabela B.2). O modelo forneceu uma ASC de 0,77 para a amostra de teste (Tabela 18) e um R^2 de 0,16.

Tabela 17 – Coeficientes estimados pela regressão logística: todos os sítios.

Variável	Coeficiente estimado	Erro Padrão	P
Interceptação	-1,140	0,917	0,2139
Etilismo [atual e ex-etilista vs nunca]	0,037	0,138	0,7917
Etilismo [atual vs ex-etilista]	0,097	0,098	0,3207
Grau de Diferenciação [moderadamente vs Bem Diferenciado]	0,146	0,092	0,1118
Grau de Diferenciação [Pouco Diferenciado vs Moderadamente e Bem Diferenciado]	0,313	0,167	0,0608
Idade	0,037	0,010	0,0003*
Idade (sim se ≥ 75 a)	0,161	0,246	0,5134
IMC	-0,028	0,024	0,2388
Linfonodomegalia	0,200	0,096	0,0380*
Odinofagia	0,131	0,098	0,1833
Perda Ponderal (%)	0,022	0,024	0,3560
Perda Ponderal (Kg)	0,042	0,039	0,2803
Perda Ponderal (S/N)	-0,130	0,127	0,3045
Sexo[M]	0,499	0,142	0,0005*
Sítio [Cavidade Oral vs Laringe]	0,168	0,107	0,1176
Sítio [Hipofaringe e Orofaringe vs Cavidade Oral e Laringe]	0,160	0,108	0,1389
Volume Clínico Ordinal cat2 vs cat1	0,112	0,161	0,4875
Volume Clínico Ordinal cat3 e cat4 vs cat1 e cat2	0,568	0,112	<0,0001*
Volume Clínico Ordinal cat4 vs cat3	0,361	0,129	0,0052*

a: anos; IMC: índice de massa corpórea; S/N: sim ou não; M: masculino

Volume clínico ordinal (cat1 $\leq$ 1,0cm³) (1,0 cm³<cat2 $\leq$ 5,0 cm³) (5,0 cm³<cat3 $\leq$ 25,0 cm³) (cat4>25 cm³)

vs: *versus* – neste caso, significando: em comparação a

* valor de *p* significativo

Tabela 18 – Curvas ROC do modelo de regressão logística para todos os sítios.

ASC: área sob a curva

O modelo de regressão logística fornece o OR de cada variável (Tabela 19).

Tabela 19 – Razão de chances de cada variável para o modelo com todos os sítios.

Variável	OR	Limite inferior r IC95%	Limite superior r IC95%
Etilismo [atual e ex-etilista <i>vs</i> nunca]	1,04	0,79	1,36
Etilismo [atual <i>vs</i> ex-etilista]	1,10	0,91	1,33
Grau de Diferenciação [moderadamente <i>vs</i> Bem Diferenciado]	1,37	1,00	1,92
Grau de Diferenciação [Pouco Diferenciado <i>vs</i> Moderadamente e Bem Diferenciado]	1,16	0,97	1,39
Idade	1,04*	1,02	1,06
Idade (sim se $\geq 75a$)	1,38	0,54	3,76
IMC	0,97	0,93	1,02
Linfonodomegalia	1,49*	1,02	2,17
Odinofagia	1,30	0,89	1,91
Perda Ponderal (%)	0,77	0,47	1,26
Perda Ponderal (Kg)	1,02	0,97	1,07
Perda Ponderal (S/N)	1,04	0,97	1,13
Sexo[M]	0,77*	0,47	1,26
Sítio [Cavidade Oral <i>vs</i> Laringe]	2,71	1,56	4,77
Sítio [Hipofaringe e Orofaringe <i>vs</i> Cavidade Oral e Laringe]	1,18	0,96	1,46
Volume Clínico Ordinal cat2 <i>vs</i> cat1	1,17	0,95	1,45
Volume Clínico Ordinal cat3 e cat4, <i>vs</i> cat1 e cat2	1,12*	0,82	1,54
Volume Clínico Ordinal cat4 <i>vs</i> cat3	1,76	1,42	2,20

a: anos; IMC: índice de massa corpórea; S/N: sim ou não; M: masculino

Volume clínico ordinal (cat1 $\leq$ 1,0cm³) (1,0 cm³<cat2 $\leq$ 5,0 cm³) (5,0 cm³<cat3 $\leq$ 25,0 cm³) (cat4>25 cm³)

IC 95%: intervalo de confiança de 95%; OR: razão de chances

vs: *versus* – neste caso, significando: em comparação a

* valor de *p* significativo

5.2.4 – Calibração dos modelos construídos para tumores de todos os sítios

Os desempenhos dos modelos de predição baseados em regressão logística e redes neurais para todos os sítios foram quantificados em termos de calibração, usando o teste Hosmer-Lemeshow "*goodness-of-fit*". De acordo com o risco previsto pelos modelos, os pacientes foram divididos em quatro categorias de risco de morte: $Q1 \leq 25\%$, $25 < Q2 \leq 50\%$, $50 < Q3 \leq 75\%$ e $Q4 > 75\%$ Tabela 20 e 21. Ambos apresentaram boa calibração (p não significativo para diferença).

Tabela 20 – Teste "*goodness-of-fit*" – modelo de regressão logística para todos os sítios.

Faixas	Grupo	média do risco calculado	mortalidade observada
0-25%	Q 1	0,19	0,20
25,1-50%	Q 2	0,39	0,37
50,01-75%	Q 3	0,64	0,60
75,01-100%	Q 4	0,85	0,83

p=0,948

Tabela 21 – Teste "*goodness-of-fit*" – modelo de redes neurais para todos os sítios.

Faixas	Grupo	média do risco calculado	mortalidade observada
0-25%	Q 1	0,00	0,00
25,1-50%	Q 2	0,36	0,34
50,01-75%	Q 3	0,64	0,62
75,01-100%	Q 4	0,85	0,84

p=0,932

5.2.5 – Comparação dos modelos construídos para todos os sítios com o TNM

Na tabela 22 encontram-se as curvas de sobrevida em cinco anos dos modelos construídos e do TNM da 7ª e 8ª edição

Tabela 22 – Curvas de sobrevida em cinco anos para tumores de todos os sítios.

Linhas pontilhadas representam intervalo de confiança de 95%

5.3 – Modelos preditivos para tumores de cavidade oral

5.3.1 – Modelos baseados em árvores

5.3.1.1 – Modelo CART para tumores de cavidade oral

Dentre os 1811 pacientes com informação de desfecho, 752 possuíam a cavidade oral como o sítio primário do tumor. Empregando todas as variáveis citadas no item 4.9, foi construído o modelo CART para cavidade oral. Para realizar validação do modelo,

usamos o DATA3 (item 4.8.3). O melhor valor de R^2 foi definido como o parâmetro para término da construção da árvore.

Analisando a figura 18, observamos a importância da espessura tumoral como parâmetro preditivo na mortalidade em cinco anos. Na subamostra cuja espessura é menor que 2,42cm, o principal definidor de mortalidade é a maior dimensão clínica. Tais informações mostram a importância do tamanho tumoral como variável prognóstica.

Na tabela 23 encontram-se informações a respeito dos testes e da sequência de divisões.

Tabela 23 – Medidas de R^2 e número de pacientes para subdivisão da amostra – modelo CART para tumores de cavidade oral.

	R^2	N	Número de divisões	AIC
Treinamento	0,111	369	3	1111,92
Validação	0,005	177		
Teste	0,083	206		

AIC: Akaike *information criterion*

R^2 : coeficiente de determinação

Figura 18 – Árvore de classificação e regressão – sítio primário cavidade oral

A tabela 24 demonstra a discriminação do modelo desenvolvido pelo método de CART, através das curvas ROC (e respectivas ASCs calculadas) das subdivisões em treinamento, validação e teste.

Tabela 24 – Curvas ROC do modelo de CART para tumores de cavidade oral.

ASC: área sob a curva

Entre todas as 27 variáveis preditivas utilizadas na construção do modelo CART, apenas três apresentaram contribuição final ao modelo (Tabela 25).

Tabela 25 – Contribuição das variáveis no modelo de CART para tumores de cavidade oral.

Variável	Número de divisões	G ²	G ²
Espessura Clínica	1	36,99	<div style="width: 100%;"></div>
Dimensão Clínica 1	1	10,67	<div style="width: 30%;"></div>
Idade (sim se ≥75a)	1	5,77	<div style="width: 15%;"></div>

5.3.1.2 – *Boosted regression trees* para tumores de cavidade oral

Empregando todas as variáveis descritas no item 3.9, construímos o modelo pelo método de BRT para tumores de cavidade oral. Utilizamos 100 camadas com número mínimo de 3 divisões e a variável data3.50.25.25 para validação.

Dentre as 27 variáveis, 17 contribuíram na elaboração final do modelo baseado em BRT. (Tabela 26). Sexo, volume clínico binário e perda ponderal categórica (com ponto de corte em 3,5kg) foram as variáveis que mais determinaram a construção deste modelo ($R^2=0,19$) e a ASC em uma subamostra para teste foi de 0,67 (Tabela 27).

Tabela 26 – Contribuição das variáveis preditivas para construção do modelo de BRT para tumores de cavidade oral.

Variável	Número de divisões	G ²	G ²
Sexo	14	5044,34	
Volume Clínico Binário	13	4454,03	
Perda Ponderal (Kg) (sim se ≥3,5kg)	7	2735,69	
Dimensão Clínica 1	3	1758,23	
Idade (sim se ≥75a)	4	1728,48	
Idade	3	1188,56	
Espessura Clínica	5	903,04	
Disfonia	7	781,21	
Grau de Diferenciação	5	710,26	
Perda Ponderal (%)	1	707,34	
Linfonodomegalia	2	652,30	
Perda Ponderal (S/N)	2	553,80	
Odinofagia	1	401,44	
Etilismo	2	332,20	
IMC	4	292,41	
Perda Ponderal (m)	1	117,31	
Dimensão Clínica 2	1	60,97	

a: anos; S/N: sim ou não; IMC: índice de massa corpórea; m: meses

Volume clínico binário (cat1≤25 cm³) (cat2>25 cm³); Volume clínico ordinal (cat1≤1,0cm³) (1,0 cm³<cat2≤5,0 cm³) (5,0 cm³<cat3≤25,0 cm³) (cat4>25 cm³)

Tabela 27 – Curvas ROC do modelo de BRT para tumores de cavidade oral.

ASC: área sob a curva

5.3.1.3 – Floresta Aleatória para tumores de cavidade oral

Elaboramos um modelo de Floresta Aleatória utilizando como parâmetro a

construção de 100 árvores com um mínimo de 10 divisões em cada árvore. Utilizamos o DATA3 como validação.

Dentre todas as 27 variáveis, a única que não contribuiu na construção de nenhuma árvore foi a presença de metástase à distância. A contribuição de todas as variáveis e discriminação do modelo encontram-se na Tabela 28 e 29. Neste modelo as três mais importantes foram volume clínico, espessura e maior dimensão clínica.

Tabela 28 - Contribuição das variáveis preditivas para construção do modelo de Floresta Aleatória para tumores de cavidade oral.

Variável	Número de divisões	G ²	G ²
Volume Clínico	77	634,08	
Espessura Clínica	79	621,16	
Dimensão Clínica 1	74	553,52	
Volume Clínico LOG	70	485,82	
Perda Ponderal (%)	62	314,03	
Volume Clínico Binário	33	307,64	
Escolaridade	58	307,38	
Idade	54	301,26	
Volume Clínico Ordinal	53	267,74	
Perda Ponderal (Kg)	54	254,90	
IMC	53	249,27	
Grau de Diferenciação	63	233,55	
Dimensão Clínica 2	40	232,12	
Perda Ponderal (Kg) (sim se ≥3,5kg)	45	200,08	
Etilismo	62	162,95	
Perda Ponderal (m)	43	143,66	
Perda Ponderal (m) (sim se ≥5m)	51	130,85	
Linfonodomegalia	45	127,11	
Sexo	39	107,34	
Raça	33	95,66	
Tabagismo	33	83,85	
Perda Ponderal (S/N)	39	76,91	
Odinofagia	39	72,19	
Idade (sim se ≥75a)	21	61,91	
Disfonia	30	61,80	

LOG: logaritmo; IMC: índice de massa corpórea; m: meses; a: anos; cat: categoria

Volume clínico ordinal (cat1≤1,0cm³) (1,0 cm³<cat2≤5,0 cm³) (5,0 cm³<cat3≤25,0 cm³) (cat4>25 cm³)

Volume clínico binário (cat1≤25 cm³) (cat2>25 cm³)

Tabela 29 – Curvas ROC do modelo de Floresta Aleatória para tumores de cavidade oral.

5.3.2 – Redes Neurais para tumores de cavidade oral

Todas as variáveis listadas no item 4.9 foram utilizadas para construção do modelo baseado em redes neurais. Para validação utilizamos o “data3.50.25.25”. Definimos três nós em uma única camada. No Anexo B encontra-se a Tabela B.3 com informações a respeito dos três grupos (treinamento, validação e teste).

Com relação à discriminação, observa-se que mesmo sendo testado em uma amostra que não foi utilizada na construção, a ASC foi de 0,75, superior ao modelo de redes neurais com todos os sítios (Tabela 30).

Tabela 30 – Curvas ROC do modelo de redes neurais para tumores de cavidade oral.

5.3.3 – Regressão Logística para tumores de cavidade oral

Na nossa subamostra com tumores de cavidade oral, 476 pacientes faleceram antes de completar o quinto ano de acompanhamento e 276 sobreviveram. Portanto é possível utilizar 27 variáveis sem comprometer a estabilidade do modelo.

Análises preliminares utilizando modelos baseados em árvores apresentaram possíveis fatores de valor preditivo e, associado ao conhecimento clínico, algumas interações factíveis. Todas as interações foram testadas, mas não foram significativas. Vários modelos de regressão logística foram construídos até chegarmos ao modelo final.

Realizamos a seleção das variáveis utilizando como método de parada o *Max Validation R²*, com direção anterógrada combinada, sendo selecionadas pelo software como variáveis com coeficientes estatisticamente significativos: idade, linfonodomegalia, sexo e volume clínico binário (com ponto de corte em 25cm³). Além dessas também entraram no modelo o grau de diferenciação e a variável e proporção da perda ponderal (Tabela 31).

Dos 752 elegíveis com sítio primário na cavidade oral, foram excluídos das análises 97 por não possuírem dados completos de todas as variáveis. Assim, 655 pacientes foram divididos em três subgrupos para treinamento (n=318), validação (n=158) e teste (n=179) (Anexo B, Tabela B.4). O modelo de regressão logística apresenta o OR de cada variável (Tabela 32).

O modelo forneceu uma ASC de 0,78 para a amostra de teste (Tabela 33) e um R² de 0,15.

Tabela 31 – Coeficientes estimados pela regressão logística: cavidade oral.

Variável	Coeficiente Estimado	Erro Padrão	P
Interceptação	-0,04	0,84	0,9624
Grau de Diferenciação [Pouco Diferenciado vs Bem Diferenciado e Moderadamente]	0,54	0,31	0,085
Idade	0,03	0,01	0,0095*
Linfonodomegalia	0,26	0,13	0,0461*
Perda Ponderal (%)	0,04	0,02	0,0905
Sexo[M]	0,48	0,17	0,0039*
Volume Clínico Binário cat1 vs cat2	-1,49	0,39	0,0001*

M: masculino; Volume clínico binário (cat1≤25 cm³) (cat2>25 cm³); vs: *versus* – neste caso, significando: em comparação a

* valor de p significativo

Tabela 32 – Razão de chances de cada variável para o modelo com tumores em cavidade oral.

Variável	OR	Limite inferior IC95%	Limite superior IC95%
Grau de Diferenciação [Pouco Diferenciado vs Bem Diferenciado e Moderadamente]	1,72	0,96	3,39
Idade	1,03*	1,01	1,06
Linfonomegalia	1,68*	1,01	2,81
Perda Ponderal (%)	1,04	1,00	1,08
Sexo[M]	2,62*	1,37	5,10
Volume Clínico Binário cat2 vs cat1	4,45*	2,16	10,03

IC 95%: intervalo de confiança de 95%; M: masculino; OR: razão de chances

Volume clínico binário (cat1 $\leq$ 25 cm³) (cat2 $>$ 25 cm³)

vs: *versus* – neste caso, significando: em comparação a

* valor de p significativo

Tabela 33 – Curvas ROC do modelo de regressão logística para tumores de cavidade oral.

ASC: área sob a curva

5.3.4 – Calibração dos modelos construídos para tumores de cavidade oral

Os desempenhos dos modelos de predição baseados em regressão logística e redes neurais para tumores de cavidade oral foram quantificados em termos de calibração, usando o teste Hosmer-Lemeshow "*goodness-of-fit*" (Tabela 34 e 35). Ambos apresentaram boa calibração ($p > 0,05$ para diferença).

Tabela 34 – Teste "*goodness-of-fit*" – modelo de regressão logística para tumores de cavidade oral.

Faixas	Grupo	média do risco calculado	mortalidade observada
0-25%	Q 1	0,21	0,38
25,1-50%	Q 2	0,42	0,39
50,01-75%	Q 3	0,61	0,57
75,01-100%	Q 4	0,87	0,86

$p=0,955$

Tabela 35 – Teste "*goodness-of-fit*" – modelo de redes neurais para tumores de cavidade oral.

Faixas	Grupo	média do risco calculado	mortalidade observada
0-25%	Q 1	0,00	0,00
25,1-50%	Q 2	0,45	0,39
50,01-75%	Q 3	0,63	0,56
75,01-100%	Q 4	0,82	0,84

$p=0,490$

5.3.5 – Comparação dos modelos construídos para tumores de cavidade oral com o TNM

Na tabela 36 encontram-se as curvas de sobrevida em cinco anos dos modelos construídos e do TNM da 7ª e 8ª edição.

Tabela 36 – Curvas de sobrevida em cinco anos para tumores de cavidade oral.

Linhas pontilhadas representam intervalo de confiança de 95%

5.4 – Modelos preditivos para tumores de orofaringe

5.4.1 – Modelos baseados em árvores

5.4.1.1 – Modelo CART para tumores de orofaringe

Empregando todas as variáveis citadas no item 4.9, foi construído o modelo

CART para todos os pacientes com tumores da orofaringe que possuíam seguimento completo. Para realizar validação do modelo, usamos o data3.50.25.25 (item 4.8.3). O melhor valor de R2 foi definido como o parâmetro para término da construção da árvore (Tabela 37; Figura 19).

Tabela 37 – Medidas de performance do modelo de CART para tumores de orofaringe.

	R ²	N	Número de divisões	Atribuições	AIC
Treinamento	0,15	212	3	7	600,99
Validação	-0,05	96			
Teste	-0,04	90			

AIC: Akaike *information criterion*

R2: coeficiente de determinação

Figura 19 – Árvore de classificação e regressão – sítio primário orofaríngeo.

Assim como identificado no modelo CART para cavidade oral, a espessura clínica também foi o principal fator preditor no modelo CART para orofaringe, entretanto com um valor de corte um pouco mais baixo. As outras variáveis que contribuíram com o modelo foram perda ponderal (%) e sexo (tabela 38).

Tabela 38 – Medidas de R² e número de pacientes para cada subdivisão da amostra – modelo de CART para tumores de orofaringe.

Variável	Número de divisões	G ²	G ²
Espessura Clínica	1	14,15	
Perda Ponderal (%)	1	11,55	
Sexo	1	6,89	

A tabela 39 demonstra a discriminação do modelo desenvolvido pelo método CART, através das curvas ROC (e respectivas ASCs calculadas) das subdivisões em treinamento, validação e teste.

Tabela 39 – Curvas ROC do modelo de CART para tumores de orofaringe.

ASC: área sob a curva

5.4.1.2 – *Boosted regression trees* para tumores de orofaringe

Empregando todas as variáveis descritas no item 3.9, construímos o modelo pelo método de BRT para tumores de orofaringe. Utilizamos 100 camadas com número mínimo de 3 divisões e a variável data3 para validação.

Dentre as 27 variáveis, 18 contribuíram na elaboração final do modelo baseado em BRT. (Tabela 40). Linfonodomegalia, sexo, idade e perda ponderal categórica (com ponto de corte em 3,5kg) foram as variáveis que mais contribuíram para a construção deste modelo ($R^2=0,10$) e a curva ROC em uma subamostra para teste foi de 0,73 (Tabela 41).

Tabela 40 – Contribuição das variáveis preditivas para construção do modelo de BRT para tumores de orofaringe.

Variável	Número de divisões	G ²	G ²
Linfonodomegalia	8	1359,59	
Sexo	9	1023,39	
Idade	4	750,63	
Perda Ponderal (Kg) (sim se ≥3,5kg)	8	473,09	
Disfonia	4	293,64	
Espessura Clínica	1	286,73	
Raça	2	283,03	
Dimensão Clínica 1	1	253,71	
Etilismo	1	63,79	
Tabagismo	2	43,24	
Dimensão Clínica 2	1	39,37	
Volume Clínico Binário	1	39,35	
Perda Ponderal (m) (sim se ≥5m)	1	27,83	
Volume Clínico Ordinal	1	26,75	
IMC	1	25,60	
Grau de Diferenciação	1	25,27	
Odinofagia	1	19,11	
Idade (sim se ≥75a)	1	5,17	

IMC: índice de massa corpórea

Volume clínico binário (cat1≤25 cm³) (cat2>25 cm³)

Volume clínico ordinal (cat1≤1,0cm³) (1,0 cm³<cat2≤5,0 cm³) (5,0 cm³<cat3≤25,0 cm³) (cat4>25 cm³)

Tabela 41 – Curvas ROC do modelo de BRT para tumores de orofaringe.

5.4.1.3 – Floresta Aleatória para tumores de orofaringe

Elaboramos um modelo de Floresta Aleatória utilizando como parâmetro a construção de 100 árvores com um mínimo de 10 divisões em cada árvore. Utilizamos o DATA3 como validação.

Dentre todas as 27 variáveis, a única que não contribuiu na construção de nenhuma árvore foi novamente a presença de metástase à distância. A contribuição de todas as variáveis e discriminação do modelo encontram-se na Tabela 42 e 43. Neste modelo as três variáveis mais importantes foram linfonodomegalia, idade e volume clínico logarítmico.

Tabela 42 - Contribuição das variáveis preditivas para construção do modelo de Floresta Aleatória para tumores de orofaringe.

Variável	Número de divisões	G ²	G ²
Linfonodomegalia	52	495,35	
Idade	56	318,21	
Volume Clínico LOG	36	216,64	
Escolaridade	37	200,58	
Etilismo	38	171,99	
Perda Ponderal (Kg)	41	157,27	
Volume Clínico	39	152,39	
IMC	40	145,72	
Perda Ponderal (%)	34	141,53	
Raça	34	127,39	
Dimensão Clínica 2	35	122,22	
Perda Ponderal (m) (sim se ≥5m)	33	109,00	
Grau de Diferenciação	31	99,98	
Dimensão Clínica 1	28	96,78	
Espessura Clínica	25	91,96	
Perda Ponderal (S/N)	24	89,21	
Odinofagia	40	87,81	
Tabagismo	26	85,30	
Perda Ponderal (Kg) (sim se ≥3,5kg)	29	80,50	
Volume Clínico Ordinal	24	76,47	
Sexo	13	46,71	
Perda Ponderal (m)	19	45,70	
Idade (sim se ≥75a)	13	42,36	
Disfonia	16	32,31	
Volume Clínico Binário	19	31,92	

LOG: logaritmo; IMC: índice de massa corpórea; m: meses; a: anos; cat: categoria

Volume clínico ordinal (cat1≤1,0cm³) (1,0 cm³<cat2≤5,0 cm³) (5,0 cm³<cat3≤25,0 cm³) (cat4>25 cm³); Volume clínico binário (cat1≤25 cm³) (cat2>25 cm³)

Tabela 43 – Curvas ROC do modelo de Floresta Aleatória para tumores de orofaringe

ASC: área sob a curva

5.4.2 – Redes Neurais para tumores de orofaringe

Todas as variáveis listadas no item 4.9, exceto “idade categórica” “perda ponderal (kg) categórica” e “perda ponderal (m) categórica”, foram utilizadas para construção do modelo para orofaringe baseado em redes neurais. Para validação utilizamos o “data3.50.25.25”. Definimos três nós na primeira camada. No Anexo B encontra-se a Tabela B.5 com informações a respeito dos três grupos (treinamento, validação e teste).

Com relação a discriminação, observa-se que mesmo sendo testado em uma amostra que não foi utilizada na construção, a ASC ainda foi de 0,65 (Tabela 44).

Tabela 44 –Curvas ROC do modelo de redes neurais para tumores de orofaringe.

ASC: área sob a curva

5.4.3 - Regressão Logística para tumores de orofaringe

Na nossa subamostra com tumores de orofaringe, 292 pacientes faleceram antes de completar o quinto ano de acompanhamento e 106 sobreviveram. Portanto é possível a construção de um modelo utilizando no máximo 10 variáveis sem comprometer a estabilidade do mesmo.

Análises preliminares utilizando modelos baseados em árvores apresentaram possíveis fatores de valor preditivo e, associado ao conhecimento clínico, algumas interações factíveis. Todas as interações foram testadas, mas não apresentaram contribuição significativa. Vários modelos de regressão logística foram construídos até chegarmos ao modelo final.

Dos 398 elegíveis para orofaringe, foram excluídos das análises 25 por não possuírem dados completos de todas as variáveis. 373 pacientes foram divididos em três grupos para treinamento (n=144), validação (n=111) e teste (n=118) (Anexo B, Tabela B.6). Realizada a seleção das variáveis utilizando como método de parada o *Max Validation R²*, direção anterógrada combinada, sendo selecionadas as seguintes variáveis com coeficientes estatisticamente significativos: idade, linfonodomegalia, sexo e volume clínico ordinal. Além delas, perda ponderal categórica (Kg) também entrou no modelo (Tabela 45). O modelo de regressão logística apresenta a razão de chances (OR) de cada variável (Tabela 46).

O modelo forneceu uma ASC de 0,72 para a amostra de teste (Tabela 47) e um R² de 0,26.

Tabela 45 – Coeficientes estimados pela regressão logística: orofaringe.

Variável	Coefficiente Estimado	Erro Padrão	p
Interceptação	-6,19	1,66	0,0002*
Idade	0,10	0,03	0,0002*
Linfonodomegalia	1,14	0,28	<,0001*
Perda Ponderal (Kg) (sim se $\geq 3,5$ kg) [Sim]	0,41	0,22	0,0672
Sexo[M]	0,92	0,31	0,0029*
Volume Clínico Ordinal cat4;cat3 vs cat2;cat1	0,55	0,26	0,0388*

M: masculino; cat: categoria; Volume clínico ordinal (cat1 $\leq$ 1,0) (1,0<cat2 $\leq$ 5,0) (5,0<cat3 $\leq$ 25,0) (cat4>25); vs: *versus* – neste caso, significando: em comparação a

* valor de p significativo

Tabela 46 – Razão de chances de cada variável para o modelo com tumores de orofaringe.

Variável	OR	Limite inferior IC95%	Limite superior IC95%
Idade	1,1*	1,05	1,17
Linfonodomegalia	9,85*	3,49	30,85
Perda Ponderal (Kg) (sim se $\geq 3,5$ kg) [Sim]	2,26	0,95	5,54
Sexo [M]	6,3*	1,93	22,4
Volume Clínico Ordinal cat4;cat3 vs cat2;cat1	1,73*	1,03	2,92

IC 95%: intervalo de confiança de 95%; OR: razão de chances

M: masculino; cat: categoria

Volume clínico ordinal (cat1 $\leq$ 1,0) (1,0<cat2 $\leq$ 5,0) (5,0<cat3 $\leq$ 25,0) (cat4>25)

vs: *versus* – neste caso, significando: em comparação a

* valor de p significativo

Tabela 47 – Curvas ROC do modelo de regressão logística para tumores de orofaringe.

ASC: área sob a curva

5.4.4 – Calibração dos modelos construídos para tumores de orofaringe

Os desempenhos dos modelos de predição baseado em regressão logística e redes neurais para tumores de orofaringe foram quantificados em termos de calibração, usando o teste Hosmer-Lemeshow "*goodness-of-fit*" (Tabela 48 e 49). Ambos apresentaram boa calibração pelo teste ($p > 0,05$ para a diferença), porém já surge algum distanciamento numérico entre os riscos esperados e observados, apesar da ausência de significância estatística.

Tabela 48 – Teste "*goodness-of-fit*" – modelo de regressão logística para tumores de orofaringe.

Faixas	Grupo	média do risco calculado	mortalidade observada
0-25%	Q 1	0,14	0,34
25,1-50%	Q 2	0,37	0,48
50,01-75%	Q 3	0,63	0,73
75,01-100%	Q 4	0,88	0,84

$p=0,189$

Tabela 49 – Teste "*goodness-of-fit*" – modelo de redes neurais para tumores de orofaringe.

Faixas	Grupo	média do risco calculado	mortalidade observada
0-25%	Q 1	0,00	0,00
25,1-50%	Q 2	0,36	0,49
50,01-75%	Q 3	0,65	0,68
75,01-100%	Q 4	0,88	0,84

$p=0,064$

5.4.5 – Comparação dos modelos construídos para tumores de orofaringe com o TNM

Na tabela 50 encontram-se as curvas de sobrevida em cinco anos dos modelos construídos e do TNM da 7ª e 8ª edição

Tabela 50 – Curvas de sobrevida em cinco anos para tumores de orofaringe.

5.5 – Modelos preditivos para tumores de hipofaringe

5.5.1 – Modelos baseados em árvores

5.5.1.1 – Modelo CART para tumores de hipofaringe

Empregando todas as variáveis citadas no item 4.9, foi construído o modelo CART para todos os 167 pacientes com tumores de hipofaringe que possuíam seguimento completo. Para realizar validação do modelo, usamos o data3.50.25.25 (item 4.8.3). O melhor valor de R^2 foi definido como o parâmetro para término da construção da árvore.

Na tabela 51 encontram-se informações a respeito dos testes.

Ao analisar o comportamento de cada variável, concluímos que o parâmetro preditivo que mais contribuiu para o modelo foi a perda ponderal (kg), seguido pelo volume tumoral (Figura 20 e tabela 52).

Tabela 51 – Medidas de R^2 e número de pacientes para cada subdivisão da amostra – modelo de CART para tumores de hipofaringe.

	R^2	N	Número de divisões	Atribuições	AIC
Treinamento	0,46	82	9	18	205,31
Validação	-0,64	44			
Teste	-0,48	41			

AIC: Akaike *information criterion*

R^2 : coeficiente de determinação

Figura 20 – Árvore de classificação e regressão – sítio primário hipofarínge.

Tabela 52 – Contribuição das variáveis no modelo de CART para tumores de hipofaringe.

Variável	Número de divisões	G ²	G ²
Perda Ponderal (Kg)	1	9,55
Volume Clínico	2	9,22
IMC	1	6,85
Disfonia	1	6,15
Raça	1	3,07
Dimensão Clínica 1	1	1,73
Etilismo	1	1,02
Grau de Diferenciação	1	0,48

IMC: índice de massa corpórea

A tabela 53 demonstra a discriminação do modelo desenvolvido pelo método de CART, através das curvas ROC (e respectivas ASCs calculadas) das subdivisões em treinamento, validação e teste.

Tabela 53 – Curvas ROC do modelo CART para tumores de hipofaringe.

ASC: área sob a curva

5.5.1.2 – *Boosted regression trees* para tumores de hipofaringe

Empregando todas as variáveis descritas no item 3.9, construímos o modelo pelo método de BRT para tumores de cavidade oral. Utilizamos 100 camadas com número mínimo de 3 divisões e a variável data3.50.25.25 para validação.

Dentre as 27 variáveis, 14 contribuíram na elaboração final do modelo baseado em BRT. (Tabela 54). Perda ponderal categórica (com ponto de corte em 3,5kg), disfonia e volume clínico binário foram as variáveis que mais determinaram a construção deste modelo ($R^2=0,557$) e a ASC em uma subamostra para teste foi de 0,73 (Tabela 55).

Tabela 54 – Contribuição das variáveis preditivas para construção do modelo de BRT para tumores de hipofaringe.

Variável	Número de divisões	G ²	G ²
Perda Ponderal (Kg) (sim se ≥3,5kg)	6	128,16
Disfonia	5	89,96
Volume Clínico Binário	7	79,70
Etilismo	5	74,99
Linfonodomegalia	3	62,05
Tabagismo	4	46,33
Perda Ponderal (S/N)	1	31,30
Dimensão Clínica 2	3	29,48
Perda Ponderal (m)	3	27,26
Idade	1	24,84
Perda Ponderal (m) (sim se ≥5m)	3	21,75
Odinofagia	1	17,29
Raça	1	10,53
Perda Ponderal (Kg)	1	3,72

Cat: categoria

Volume clínico binário (cat1≤25 cm³) (cat2>25 cm³)

Tabela 55 – Curvas ROC do modelo de BRT para tumores de hipofaringe.

ASC: área sob a curva

5.5.1.3 – Floresta Aleatória para tumores de hipofaringe

Elaboramos um modelo de Floresta Aleatória utilizando como parâmetro a construção de 100 árvores com um mínimo de 10 divisões em cada árvore. Utilizamos o data3.50.25.25 como validação.

Dentre todas as 27 variáveis, as únicas que não contribuíram na construção de nenhuma árvore foi novamente a presença de metástase à distância, idade categórica (ponto de corte em 75 anos) e sexo. A contribuição de todas as variáveis e discriminação do modelo encontram-se na Tabela 56 e 57. Neste modelo as três mais importantes foram perda ponderal, raça e idade.

Tabela 56 - Contribuição das variáveis preditivas para construção do modelo de Floresta Aleatória para tumores de hipofaringe.

Variável	Número de divisões	G ²	G ²
Perda Ponderal (Kg)	22	117,57	
Raça	17	111,56	
Idade	13	67,08	
Volume Clínico LOG	8	64,05	
IMC	14	62,51	
Espessura Clínica	16	51,68	
Perda Ponderal (%)	10	51,51	
Disfonia	14	48,45	
Volume Clínico	10	37,39	
Dimensão Clínica 2	9	30,42	
Perda Ponderal (Kg) (sim se ≥3,5kg)	12	28,90	
Escolaridade	8	25,94	
Volume Clínico Binário	9	25,50	
Grau de Diferenciação	7	25,09	
Volume Clínico Ordinal	11	21,17	
Perda Ponderal (m) (sim se ≥5m)	9	20,48	
Linfonodomegalia	7	20,09	
Etilismo	13	16,94	
Perda Ponderal (S/N)	8	16,76	
Tabagismo	6	15,71	
Perda Ponderal (m)	6	13,60	
Odinofagia	7	10,98	
Dimensão Clínica 1	4	8,63	

LOG: logaritmo; IMC: índice de massa corpórea; m: meses; a: anos; cat: categoria;

S/N: sim ou não

Volume clínico binário (cat1≤25 cm³) (cat2>25 cm³)

Volume clínico ordinal (cat1≤1,0cm³) (1,0 cm³<cat2≤5,0 cm³) (5,0 cm³<cat3≤25,0 cm³) (cat4>25 cm³)

Tabela 57 – Curvas ROC do modelo de Floresta Aleatória para tumores de hipofaringe.

ASC: área sob a curva

5.5.2 – Redes Neurais para tumores de hipofaringe

Todas as variáveis listadas no item 4.9, exceto “idade categórica” “perda ponderal (kg) categórica” e “perda ponderal (m) categórica”, foram utilizadas para construção do modelo para hipofaringe baseado em redes neurais. Para validação utilizamos o “data3.50.25.25”. Definimos três nós na primeira camada. No Anexo B encontra-se a Tabela B.7 com informações a respeito dos três grupos (treinamento, validação e teste)

Com relação à discriminação, seguem abaixo as curvas ROC das três subamostras utilizadas (Tabela 58).

Tabela 58 – Curvas ROC do modelo de redes neurais para tumores de hipofaringe.

ASC: área sob a curva

5.5.3 – Regressão Logística para tumores de hipofaringe

Na nossa subamostra com tumores de hipofaringe, 132 pacientes faleceram antes de completar o quinto ano de acompanhamento e 35 sobreviveram. Portanto é possível a construção de um modelo utilizando no máximo 3 variáveis sem comprometer a estabilidade do modelo.

Análises preliminares utilizando modelos baseados em árvores apresentaram possíveis fatores de valor preditivo e, associado ao conhecimento clínico, algumas interações factíveis. Todas as interações foram testadas, mas não foram significativas. Vários modelos de regressão logística foram construídos até chegarmos a este modelo final.

Realizada a seleção das variáveis utilizando como método de parada o *Max Validation* R^2 , direção anterógrada combinada, sendo selecionadas pelo *software* como variáveis com coeficientes estatisticamente significativos: idade e linfonodomegalia. Além desses dois fatores, etilismo também entrou no modelo (Tabela 59).

Dos 167 elegíveis com sítio primário na hipofaringe, foram excluídos das análises 17 por não possuírem dados completos de todas as variáveis. 150 pacientes foram divididos em três subgrupos para treinamento (n=47), validação (n=56) e teste (n=47) (Anexo B, Tabela B.8). O modelo de regressão logística fornece a razão de chances (*Odds ratio*) de cada variável (Tabela 60).

O modelo forneceu uma ASC de 0,45 para a amostra de teste (Tabela 61) e um R^2 de -1,3.

Tabela 59 – Coeficientes estimados pela regressão logística: hipofaringe.

Variável	Coeficiente Estimado	Erro Padrão	p
Interceptação	16	3440	0,9963
Etilismo [etilista]	-5	3440	0,9988
Etilismo [ex-etilista]	-8	3440	0,9982
Idade	0	0	0,0399*
Linfonodomegalia	1	1	0,0338*

* valor de p significativo

Tabela 60 – Razão de chances de cada variável para o modelo com tumores de hipofaringe.

Variável	OR	Limite inferior IC95%	Limite superior IC95%
Etilismo[etilista]	12,21	1,30	362,89
Etilismo[ex-etilista]	0,00	0,00	1,26
Idade	0,87*	0,75	0,98
Linfonomegalia	15,72*	1,47	307,50

IC 95%: intervalo de confiança de 95%; OR: razão de chances

* valor de p significativo

Tabela 61 – Curvas ROC do modelo de regressão logística para tumores de hipofaringe.

ASC: área sob a curva

5.5.4 – Calibração dos modelos construídos para tumores de hipofaringe

Os desempenhos dos modelos de predição baseado em regressão logística e redes neurais para tumores de hipofaringe foram quantificados em termos de calibração, usando o teste Hosmer-Lemeshow "*goodness-of-fit*" (Tabela 62 e 63). O modelo realizado por regressão logística não se apresentou calibrado ($p < 0,000001$), ao contrário do modelo por redes neurais.

Tabela 62 – Teste "*goodness-of-fit*" – modelo de regressão logística para tumores de hipofaringe.

Faixas	Grupo	média do risco calculado	mortalidade observada
0-25%	Q 1	0,13	0,75
25,1-50%	Q 2	0,38	0,63
50,01-75%	Q 3	0,61	0,68
75,01-100%	Q 4	0,93	0,81

$p < 0,000001$

Tabela 63 – Teste "*goodness-of-fit*" – modelo de redes neurais tumores de hipofaringe.

Faixas	Grupo	média do risco calculado	mortalidade observada
0-25%	Q 1	0,00	0,00
25,1-50%	Q 2	0,49	0,50
50,01-75%	Q 3	0,61	0,79
75,01-100%	Q 4	0,86	0,80

p=0,067

5.5.5 – Comparação dos modelos construídos para tumores de hipofaringe com o TNM

Na tabela 64 encontram-se as curvas de sobrevida em cinco anos dos modelos construídos e do TNM da 7ª e 8ª edição

Tabela 64 – Curvas de sobrevida em cinco anos para tumores de hipofaringe.

Linhas pontilhadas representam intervalo de confiança de 95%

5.6 – Modelos preditivos para tumores de laringe

5.6.1 – Modelos baseados em árvores

5.6.1.1 – Modelo CART para tumores de laringe

Empregando todas as variáveis citadas no item 4.9, foi construído o modelo CART para todos os 494 pacientes com seguimento completo. Para realizar validação do modelo, usamos o data3.50.25.25 (item 4.8.3). O melhor valor de R^2 foi definido como o parâmetro para término da construção da árvore (Tabela 65 e Figura 21).

Tabela 65 – Medidas de R^2 e número de pacientes para cada subdivisão da amostra - modelo de CART para tumores de laringe.

	R^2	N	Número de divisões	Atribuições	AIC
Treinamento	0,160	242	5	23	734,936
Validação	0,054	118			
Teste	0,023	134			

AIC: Akaike *information criterion*

R^2 : coeficiente de determinação

Figura 21 – Árvore de classificação e regressão – sítio primário laringe.

Em relação aos tumores cujo sítio primário foi laringe, o volume tumoral clínico foi o principal fator preditor no modelo CART, possuindo como ponto de corte o volume aproximado de 2,34cm³. As outras variáveis que contribuíram com o modelo foram perda ponderal (%) e sexo (tabela 66).

Tabela 66 – Contribuição das variáveis no modelo de CART para tumores de laringe.

Variável	Número de divisões	G ²	G ²
Volume Clínico	1	23,31
Perda Ponderal (%)	1	15,88
Idade	1	8,21
Grau de Diferenciação	1	4,63
Disfonia	1	2,92

A tabela 67 demonstra a discriminação do modelo desenvolvido pelo método de CART, através das curvas ROC (e respectivas ASCs calculadas) das subdivisões em treinamento, validação e teste.

Tabela 67 – curvas ROC do modelo de CART para tumores de laringe

ASC: área sob a curva

5.6.1.2 – *Boosted regression trees* para tumores de laringe

Empregando todas as variáveis descritas no item 3.9, construímos o modelo pelo método de BRT para tumores de laringe. Utilizamos 100 camadas com número mínimo de 3 divisões e a variável data3 para validação.

Dentre as 27 variáveis, 18 contribuíram na elaboração final do modelo baseado em BRT. (Tabela 68). Grau de diferenciação, linfonodomegalia e segunda dimensão clínica foram as variáveis que mais determinaram a construção deste modelo ($R^2=0,15$) e a ASC em uma subamostra para teste foi de 0,70 (Tabela 69).

Tabela 68 – Contribuição das variáveis preditivas para construção do modelo de BRT para tumores de laringe.

Variável	Número de divisões	G ²	G ²
Grau de Diferenciação	11	3129,21	
Linfonodomegalia	9	2195,10	
Dimensão Clínica 2	7	2177,62	
Volume Clínico Binário	13	1811,47	
Perda Ponderal (S/N)	7	1027,19	
Volume Clínico Ordinal	3	1026,90	
Odinofagia	13	876,30	
Disfonia	10	851,79	
Idade (sim se ≥75a)	3	841,24	
Volume Clínico	2	833,04	
Espessura Clínica	2	494,69	
Idade	4	459,48	
Perda Ponderal (%)	3	272,42	
Etilismo	9	248,03	
Perda Ponderal (Kg) (sim se ≥3,5kg)	2	191,74	
Sexo	12	189,22	
Tabagismo	1	82,69	
Perda Ponderal (Kg)	1	8,57	

IMC: índice de massa corpórea; cat=categoria

Volume clínico binário (cat1≤25 cm³) (cat2>25 cm³)

Volume clínico ordinal (cat1≤1,0cm³) (1,0 cm³<cat2≤5,0 cm³) (5,0 cm³<cat3≤25,0 cm³) (cat4>25 cm³)

Tabela 69 – Curvas ROC do modelo de BRT para tumores de laringe.

ASC: área sob a curva

5.6.1.3 – Floresta Aleatória para tumores de laringe

Elaboramos um modelo de Floresta Aleatória utilizando como parâmetro a

construção de 100 árvores com um mínimo de 10 divisões em cada árvore. Utilizamos o data3.50.25.25 como validação.

Dentre todas as 27 variáveis, a única que não contribuiu na construção de nenhuma árvore foi a presença de metástase à distância. A contribuição de todas as variáveis e discriminação do modelo encontram-se na Tabela 70 e 71. Neste modelo as três mais importantes foram as duas maiores dimensões clínicas e volume clínico.

Tabela 70 - Contribuição das variáveis preditivas para construção do modelo de Floresta Aleatória para tumores de laringe.

Variável	Número de divisões	G ²	G ²
Dimensão Clínica 2	87	621,27	
Volume Clínico	81	534,03	
Dimensão Clínica 1	81	528,99	
Idade	83	432,55	
Volume Clínico LOG	64	426,36	
Grau de Diferenciação	76	362,00	
Espessura Clínica	72	345,05	
Escolaridade	69	340,74	
Perda Ponderal (Kg)	69	327,77	
Volume Clínico Ordinal	64	252,53	
IMC	61	247,99	
Etilismo	66	223,97	
Linfonodomegalia	64	196,91	
Perda Ponderal (%)	39	171,72	
Perda Ponderal (m) (sim se ≥5m)	55	166,76	
Perda Ponderal (S/N)	43	145,52	
Odinofagia	46	138,54	
Perda Ponderal (m)	43	133,08	
Tabagismo	45	117,29	
Raça	39	116,86	
Perda Ponderal (Kg) (sim se ≥3,5kg)	39	110,91	
Sexo	27	90,20	
Disfonia	30	85,95	
Volume Clínico Binário	16	28,28	
Idade (sim se ≥75a)	6	16,64	

LOG: logaritmo; IMC: índice de massa corpórea; m: meses; a: anos; cat: categoria

Volume clínico binário (cat1 ≤ 25 cm³) (cat2 > 25 cm³)

Volume clínico ordinal (cat1 ≤ 1,0 cm³) (1,0 cm³ < cat2 ≤ 5,0 cm³) (5,0 cm³ < cat3 ≤ 25,0 cm³) (cat4 > 25 cm³)

Tabela 71 – Curvas ROC do modelo de Floresta Aleatória para tumores de laringe.

ASC: área sob a curva

5.6.2 – Redes Neurais para tumores de laringe

Todas as variáveis listadas no item 4.9, exceto “idade categórica”, “perda ponderal (kg) categórica” e “perda ponderal (m) categórica”, foram utilizadas para construção do modelo para laringe baseado em redes neurais. Para validação utilizamos o “data3.50.25.25”. Definimos três nós na primeira camada. No Anexo B encontra-se a Tabela B.9 com informações a respeito dos três grupos (treinamento, validação e teste).

Com relação a discriminação, observa-se que mesmo sendo testado em uma amostra que não foi utilizada na construção, a ASC foi de 0,75. (Tabela 72).

Tabela 72 – Curvas ROC do modelo de redes neurais para tumores de laringe.

ASC: área sob a curva

5.6.3 - Regressão Logística para tumores de laringe

Na nossa subamostra com tumores de laringe, 263 pacientes faleceram antes de completar o quinto ano de acompanhamento e 231 sobreviveram. Portanto é possível utilizar 23 variáveis sem comprometer a estabilidade do modelo.

Análises preliminares utilizando modelos baseados em árvores apresentaram possíveis fatores de valor preditivo e, associado ao conhecimento clínico, algumas interações factíveis. Todas as interações foram testadas, mas não foram significativas. Vários modelos de regressão logística foram construídos até chegarmos ao modelo final.

Realizamos a seleção das variáveis utilizando como método de parada o *Max Validation* R^2 , com direção anterógrada combinada, sendo selecionadas pelo software como variáveis com coeficientes estatisticamente significativos: idade, linfonodomegalia, e volume clínico binário LA (com ponto de corte em 3cm^3). Além dessas também entraram no modelo o grau de diferenciação e a perda ponderal S/N (Tabela 73).

Dos 494 elegíveis com sítio primário na laringe, foram excluídos das análises 71 por não possuírem dados completos de todas as variáveis. Assim, 423 pacientes foram divididos em três subgrupos para treinamento ($n=158$), validação ($n=143$) e teste ($n=122$) (Anexo B, Tabela B.10). O modelo de regressão logística apresenta o OR de cada variável (Tabela 74).

O modelo forneceu uma ASC de 0,71 para a amostra de teste (Tabela 75) e um R^2 de 0,10.

Tabela 73 – Coeficientes estimados pela regressão logística: laringe.

Variável	Coeficiente Estimado	Erro Padrão	P
Interceptação	-4,67	1,45	0,0013*
Grau de Diferenciação [Pouco Diferenciado e Moderadamente vs Bem Diferenciado]	0,19	0,19	0,3218
Idade	0,04	0,02	0,0256*
Linfonodomegalia	0,43	0,22	0,0477*
Perda Ponderal (S/N)	0,22	0,20	0,2575
Volume Clínico Binário LA	1,54	0,40	0,0001*

cat: categoria; vs: *versus* – neste caso, significando: em comparação a Volume clínico binário LA (cat1 $\leq 3\text{ cm}^3$) (cat2 $>3\text{ cm}^3$)

* valor de p significativo

Tabela 74 – Razão de chances de cada variável para o modelo com tumores de laringe.

Variável	OR	Limite inferior IC95%	Limite superior IC95%
Grau de Diferenciação [Pouco Diferenciado e Moderadamente vs. Bem Diferenciado]	1,21	0,83	1,77
Idade	1,05*	1,01	1,09
Linfonodomegalia	2,36*	1,02	5,65
Perda Ponderal (S/N)	1,56	0,72	3,40
Volume Clínico Binário LA	4,66*	2,15	10,55

IC 95%: intervalo de confiança de 95%.cat: categoria; S/N: sim ou não

OR: razão de chances; Volume clínico binário LA (cat1 $\leq$ 3 cm³) (cat2>3 cm³)

vs: *versus* – neste caso, significando: em comparação a

* valor de p significativo

Tabela 75 – Curvas ROC do modelo de regressão logística para tumores de laringe.

ASC: área sob a curva

5.6.4 – Calibração dos modelos construídos para tumores de laringe

Os desempenhos dos modelos de predição baseado em regressão logística e redes neurais para tumores de laringe foram quantificados em termos de calibração, usando o teste Hosmer-Lemeshow "*goodness-of-fit*" (Tabela 76 e 77). Ambos os modelos estavam bem calibrados ($p > 0,05$ para a diferença).

Tabela 76 – Teste "*goodness-of-fit*" – modelo de regressão logística para tumores de laringe.

Faixas	Grupo	média do risco calculado	mortalidade observada
0-25%	Q 1	0,19	0,17
25,1-50%	Q 2	0,36	0,41
50,01-75%	Q 3	0,65	0,57
75,01-100%	Q 4	0,82	0,70

$p=0,302$

Tabela 77 – Teste "*goodness-of-fit*" – modelo de redes neurais para tumores de laringe.

Faixas	Grupo	média do risco calculado	mortalidade observada
0-25%	Q 1	0,20	0,16
25,1-50%	Q 2	0,37	0,30
50,01-75%	Q 3	0,64	0,60
75,01-100%	Q 4	0,80	0,90

$p=0,328$

5.6.5 – Comparação dos modelos construídos para tumores de laringe com o TNM

Na tabela 78 encontram-se as curvas de sobrevida em cinco anos dos modelos construídos e do TNM da 7ª e 8ª edição

Tabela 78 – Curvas de sobrevida em cinco anos para tumores de laringe.

Linhas pontilhadas representam intervalo de confiança de 95%

5.7 - Incremento da performance dos modelos preditivos após adição do status HPV

5.7.1 – Tumores de todos os sítios

Dentre os 1811 pacientes incluídos nas análises, 692 pacientes possuíam informações de HPV, sendo que 411 pacientes faleceram antes de completar o quinto ano de acompanhamento e 281 sobreviveram. Portanto é possível a construção de um modelo de regressão utilizando no máximo 28 variáveis sem comprometer a estabilidade do modelo. Criamos dois modelos de regressão para esta subamostra. No primeiro incluímos apenas as variáveis selecionadas durante a construção do modelo de regressão logística para todos os sítios e no segundo incluímos a entrada do HPV no modelo. Utilizamos a seleção passo a passo para selecionar variáveis, porém nenhum dos dois métodos (AIC ou *Max Validation R²*) incorporaram o HPV ao modelo. Para avaliar seu poder preditivo, forçamos a entrada da variável HPV ao modelo. Contudo, nas análises, observamos que

não houve melhora significativa ao adicionar o HPV em performance do modelo, com melhora do R^2 e ASC nas subamostras de treinamento, porém, piora das mesmas medidas nas subamostras de teste (Tabela 79 e 80)

Tabela 79 – Comparação dos modelos de regressão logística após a inclusão do HPV como variável independente na subamostras de teste.

	Sem HPV	Com HPV
R^2	0,1744	0,1584
ASC	0,724	0,717

ASC: área sob a curva

Tabela 80 – Parâmetros obtidos quando o HPV foi incluído no modelo de regressão para todos os sítios.

	Coefficiente estimado	Erro padrão	OR	Limite inferior IC95%	Limite superior IC95%	p
HPV	-0,31	0,28	1,87	0,66	5,93	0,2578

OR: razão de chances

IC 95%: intervalo de confiança de 95%

HPV: papilomavírus humano

5.7.2 – Tumores de orofaringe

Dentre os 398 pacientes com tumores de orofaringe incluídos nas análises, 119 pacientes possuíam informações de HPV, sendo que 77 pacientes faleceram antes de completar o quinto ano de acompanhamento e 44 sobreviveram. Portanto é possível a construção de um modelo de regressão utilizando no máximo 4 variáveis sem comprometer a estabilidade do modelo. Criamos dois modelos de regressão para esta subamostra. No primeiro incluímos apenas as variáveis selecionadas durante a construção do modelo de regressão logística para tumores de orofaringe e no segundo incluímos a entrada do HPV no modelo. Utilizamos a seleção passo a passo para selecionar variáveis, porém nenhum dos dois métodos (AIC ou *Max Validation R²*) incorporaram o HPV ao modelo. Para avaliar seu poder preditivo, forçamos a entrada da variável HPV ao modelo.

Contudo, nas análises, observamos que não houve melhora significativa ao adicionar o HPV em performance do modelo (Tabela 81 e 82).

Tabela 81 – Comparação dos modelos de regressão logística após a inclusão do HPV como variável independente na subamostras de teste.

	Sem HPV	Com HPV
R²	-0,374	-0,278
ASC	0,84	0,84

ASC: área sob a curva

Tabela 82 – Parâmetros obtidos quando o HPV foi incluído no modelo de regressão para tumores de orofaringe

	Coefficiente estimado	Erro padrão	OR	Limite inferior IC95%	Limite superior IC95%	p
HPV	-0,71	0,59	0,24	0,02	2,59	0,2342

OR: razão de chances

IC 95%: intervalo de confiança de 95%

HPV: papilomavírus humano

DISCUSSÃO

6.0 – DISCUSSÃO

A busca ativa por modelos preditivos é um grande objetivo em várias áreas do conhecimento humano, como economia, ecologia, sociologia etc. Ela vem ganhando força com o desenvolvimento de novos métodos estatísticos, bem como o melhor entendimento de como estas ferramentas funcionam, especialmente no contexto de saúde, com seus aspectos únicos. A identificação de pacientes com câncer com pior prognóstico é um dos grandes desafios para o sucesso no tratamento oncológico, seja ele a cura ou o aumento da sobrevida.

A estimativa precisa e individualizada da sobrevida em pacientes com CECP, sem dúvida, levará a uma melhora nas estratégias terapêuticas e de cuidados, minimizando os riscos de subtratamento ou excesso do mesmo. O atual sistema TNM se mostra limitado neste sentido. Por isso, mineramos dados e desenvolvemos modelos para prever o resultado com base nas características demográficas e clínicas dos pacientes antes do tratamento. Enquanto médicos focam no tratamento ideal, pacientes estão interessados no prognóstico e o TNM não consegue responder com precisão nem a uma pergunta nem a outra, especialmente se este paciente não corresponde ao perfil de pacientes geralmente incluídos em estudos internacionais para desenvolvimento de modelos preditivos. Até o momento, desconhecemos estudos nacionais que criaram uma ferramenta preditiva de mortalidade em cinco anos para câncer de cabeça e pescoço.

Um dos principais objetivos do projeto GENCAPO foi a busca por fatores ambientais, clínicos, histopatológicos e moleculares associados ao desenvolvimento e prognóstico de CECP, bem como identificar marcadores de agressividade. Foi graças ao esforço por 15 anos de inúmeros entrevistadores, pesquisadores e médicos, bem como a inestimável contribuição dos pacientes, que este estudo foi possível.

Para atingirmos nosso objetivo, selecionamos em meio a um vasto banco de dados todas as variáveis demográficas e clínicas disponíveis que pudessem ter alguma relação com sobrevida. Da mesma forma, para maior abrangência populacional, decidimos manter os pacientes provenientes dos cinco centros que fazem parte do GENCAPO, mesmo sabendo que alguns centros incluíram e acompanharam poucos pacientes.

Após a compilação de todos os dados provenientes do servidor central do GENCAPO, identificamos dados faltantes de desfecho em 25,8% da nossa amostra.

Foram checados todos os questionários do GENCAPO em busca de tais informações. Investigamos também prontuários físicos e eletrônicos no Instituto Arnaldo Vieira de Carvalho, Hospital Heliópolis, Instituto do Câncer do Estado de São Paulo (ICESP) e HC-FMUSP, não apenas para obter dados faltantes como também para checar inconsistências dos dados. Em seguida conferimos o Registro de Câncer do ICESP e ferramentas *online* como: obituário de jornais, cadastro nacional de falecidos (<https://www.falecidosnobrasil.org.br/>), redes sociais e livros de registro de sepultamentos em diversos cemitérios.

Ao analisar a população do nosso estudo, observamos que mesmo com a exclusão de 13,4% dos pacientes devido à perda de acompanhamento antes de completar cinco anos, não houve alteração estatisticamente significativa no perfil dos pacientes, portanto, acreditamos que foi aceitável a análise apenas dos dados de desfecho disponíveis(121).

Após a compilação das informações obtidas, alguns fatores preditores permaneceram com dados faltantes. A simples exclusão desses pacientes não era uma alternativa adequada, pois reduziria bastante o tamanho da amostra, com grandes chances de geração de vieses e perda na acurácia da ferramenta preditiva obtida. Respaldados pela literatura(76, 77), imputamos dados relacionados às três dimensões clínicas e tempo de perda de peso.

Inicialmente os dados do sítio primário apresentavam 31 pacientes com categoria dúbia: cavidade oral/orofaringe. Segundo Tierney *et al*(122), é possível compreender e resolver este tipo de dilema através do modelo de CART. Utilizando esta ferramenta, definindo esta categoria como “orofaringe”.

O perfil demográfico do paciente incluído neste estudo é de pessoas do sexo masculino, na sexta década de vida, com escolaridade até o quinto ano do ensino fundamental, tabagistas e etilistas, tratadas pelo Sistema Único de Saúde. O sítio do tumor primário mais frequente foi cavidade oral e mais de 60% dos pacientes possuíam tumores avançados no momento da entrevista. Estes dados refletem a realidade da maioria dos pacientes com CECp no Brasil(2).

Com o objetivo de prever melhor a evolução destes pacientes, iniciamos a mineração de dados utilizando toda a população incluída, independentemente do sítio tumoral primário. Apesar de modelos sítio-específicos serem o padrão para predição de

mortalidade, optamos por tentar elaborar um modelo para todos os sítios, aproveitando o tamanho amostral disponível. Nossa estratégia se provou acertada, pois conseguimos desenvolver um modelo com boa discriminação e calibração, claramente superior ao TNM.

As análises de exploração através de modelos baseados em árvores e o modelo de regressão logística para tumores de todos os sítios mostraram como principais preditores: idade, linfonodomegalia, sexo e volume clínico ordinal. Além disto estas análises apontaram para possível categorização da variável “volume clínico” e linfonodomegalia.

Apesar de não rotineiramente coletado no projeto GENCAPO, optamos por explorar o volume tumoral, dado frequentemente usado em estudos de resposta do CECP à radioterapia, como variável preditiva de morte. Achamos esta variável atraente, pois compreende as três dimensões tumorais em apenas um número, o que auxilia na desejável diminuição do tamanho do modelo. Apesar de certa discordância entre estudos acerca da fórmula usada, optamos pela fórmula de Monga *et al*, por ser prática e possuímos as três variáveis para o cálculo (altura, largura e profundidade).

Através dos modelos baseados em árvores foi possível categorizar volume clínico de duas maneiras. Primeiro, chamada de “volume clínico binário”, utilizamos apenas um ponto de corte em: $\leq 25\text{cm}^3$ e $> 25\text{cm}^3$; e segundo, chamada de “volume clínico ordinal”, utilizamos três pontos de corte: $\leq 1\text{cm}^3$; $1\text{cm}^3 < x \leq 5\text{cm}^3$; $> 5\text{cm}^3 < x \leq 25\text{cm}^3$; $> 25\text{cm}^3$. Tal categorização foi extremamente útil nas nossas análises por entendemos que os pontos fora da curva observados na distribuição desta variável poderiam comprometer as construções dos modelos. Além disso, a categorização deve se mostrar mais útil em futura aplicação clínica deste modelo, já que muitas das medidas clínicas não podem ser precisamente mensuradas, porém sua ordem de grandeza, sim.

Em CECP o volume tumoral pode se correlacionar melhor com o prognóstico do que o estadiamento tumoral convencional. Alguns estudos publicados nos últimos anos sugerem a inclusão do volume tumoral no sistema TNM. Em uma metanálise incluindo 15 estudos e 2447 pacientes, Xie *et al*(123) sugerem que o volume tumoral é um importante preditor de mortalidade para pacientes com CECP utilizando como ponto de corte $24,48\text{cm}^3$. Em 2017, estudo realizado por Dejacó *et al* mostraram que o volume tumoral teve impacto significativo na sobrevida global de paciente com CECP. Utilizando

regressão de Cox, o volume tumoral foi a única covariável com achados significativos com um aumento de 1,4% no risco de morte a cada cm^3 de aumento do volume tumoral (95% IC 0,8%-2,0%; $p < 0,001$)(124).

No nosso modelo de regressão logística para todos os sítios, o volume clínico estava categorizado em quatro níveis. Os pacientes portadores de tumores maiores que 5cm^3 possuíam um risco de morte, ajustado pelos outros fatores de risco, 76,4% maior quando comparados com tumores menores.

Cerca de um terço dos linfonodos do corpo localizam-se na cabeça e pescoço. Quando acometidos pela neoplasia, são um importante fator preditivo para pacientes com CECP(125). Os modelos classicamente usados graduam esta informação pela dimensão e pelo número de linfonodos comprometidos. Optamos pela simples dicotomização em negativos e positivos por alguns motivos: primeiramente, não dispúnhamos do tamanho nem do número de linfonodos clinicamente acometidos, apenas do seu estágio clínico N. Deste modo, não consideramos que a comparação com o TNM seria justa se usássemos a própria informação do status N do paciente como preditor em nosso modelo. Por fim, algumas análises informais, não descritas nos métodos deste estudo, demonstraram que o principal ponto de corte para discriminação de risco conforme o status N era de fato a divisão entre cN0 e cN1/2/3. Ao final, esta variável apareceu como importante preditor em todos os modelos de regressão, sendo a variável que mais contribuiu no modelo baseado na Floresta Aleatória e o melhor discriminador no CART após volume clínico tumoral.

No modelo multivariado da regressão logística o fato do paciente possuir linfonodomegalia aumentou o risco de morte em cinco anos em 49% ($p=0,038$; IC95% 2-117), tal impacto está em concordância com a literatura. Segundo Cho *et al*, à medida que as células tumorais se disseminam para linfonodos regionais, a doença vai ganhando capacidade de promover metástases à distância(126).

Já o sexo do paciente como fator preditivo de mortalidade não está bem estabelecido na literatura. Utilizando análises univariadas, alguns estudos apontam para um prognóstico muito melhor em mulheres, mesmo ajustando por variáveis de confusão(127), enquanto outros, utilizando modelos multivariados, não demonstraram nenhuma diferença entre os sexos(128). Nossas análises foram realizadas através de

modelos multivariados, pois entendemos que análises univariadas para construção de modelos preditivos podem levar a má compreensão do poder de determinada variável em relação ao desfecho.

As análises utilizando métodos baseados em árvores mostraram a importância do sexo como preditor. No CART este parâmetro foi o principal discriminador de morte no subgrupo composto por tumores com volumes clínicos menores que 7cm³. No BRT foi a variável que mais contribuiu para a construção deste modelo.

Na regressão logística o fato do paciente ser do sexo masculino aumentou o risco de morte em 71% ($p < 0,004$; IC95% 56-77). Este foi um dos achados que mais chamou a nossa atenção, pois mesmo sendo uma amostra composta por pacientes do sexo masculino na grande maioria, havia uma quantidade suficiente de mulheres em todos os subgrupos a fim de se encontrar essa relação. Esta é uma variável que é facilmente obtida e, tudo indica, pouco explorada como preditor robusto de morte. Vale ressaltar que, como em vários outros estudos, a variável sexo acaba sendo usada para controlar algumas variáveis de confusão desconhecidas. Em estudos com algum tipo de pareamento, é frequente o pareamento por sexo, mesmo que não haja sinais de causalidade com a doença. Vale ainda lembrar que, independente do mecanismo (causal ou não) por trás da associação entre sexo e morte, seu valor preditivo foi inegável nesta amostra de pacientes.

Assim como o sexo, outra variável com dados facilmente disponíveis e que foi muito importante em nosso modelo foi a idade. Alguns estudos sugerem que a idade é uma variável preditiva de morte tão forte quanto outros fatores conhecidos na literatura, principalmente quando categorizada. Medow *et al* (129) sugeriu a dicotomização da idade e utilizou como ponto de corte 66,5 anos. Nossas análises apontaram para o ponto de corte simplificado em 75 anos, com o qual estabelecemos duas categorias. No modelo de regressão, a idade foi uma variável estatisticamente significativa e com aumento no risco de óbito de 3,8% por ano de vida ($p < 0,0003$; IC95% 1,8-5,3).

Assim, estabelecemos um modelo simples, com variáveis de dados facilmente coletáveis na prática clínica, para todos os sítios da cabeça e pescoço. Apesar da variável “sítio” ter sido usada no modelo, ela não foi selecionada como significativa. Isso demonstrou o quanto os CECP de diferentes sítios têm em comum. Contudo, as particularidades de cada subsítio ainda eram um campo a se explorar.

O subsítio com maior número de pacientes era o de tumores de cavidade oral e o modelo de regressão logística para este sítio apresentou como variáveis estatisticamente significantes: idade (OR=1,03; IC 95% 1,01-1,06), linfonodomegalia (OR=1,7; IC 95% 1,01-2,8), sexo (masculino, OR=2,6; IC 95% 1,4-5,1) e volume clínico binário (OR=4,4; IC 95% 2,2-10,0). Em comparação com o modelo de regressão logística para tumores de todos os sítios, a única diferença em relação as variáveis estatisticamente significantes foi a categorização do volume clínico de quatro para duas categorias.

As análises preliminares para construção dos modelos de cavidade oral mostraram a importância da espessura tumoral como fator preditivo. Utilizando o modelo CART, a espessura tumoral foi variável que mais contribuiu para a construção deste modelo e o ponto de corte encontrado foi 24,2mm. Em nossa amostra a mediana da espessura tumoral de 19,4mm. Existe diferença conceitual entre espessura tumoral e profundidade de invasão (DOI). Embora as definições sejam semelhantes, o primeiro se refere à terceira dimensão tumoral, já o segundo, à extensão tumoral profunda a partir da base do tumor (superfície epitelial). A DOI juntamente com a proliferação microvascular causada pelo crescimento neoplásico podem levar a aproximação tumoral de vasos sanguíneos e linfáticos, facilitando o processo metastático(130). Em 2014 foi publicado um estudo multicêntrico no qual foi sugerida a inclusão da DOI ao TNM utilizando como pontos de corte 5mm e 10mm(131).

Um subsítio que tem se isolado no novo sistema TNM é a orofaringe, graças ao crescente número de tumores causados pelo HPV. Nossa positividade para o HPV em câncer de orofaringe foi de 11,8%. Assim, não esperávamos modelos muito distintos daqueles desenvolvidos para tumores de cavidade oral. Nas análises utilizando os modelos baseados em árvores para tumores em orofaringe, o preditor que mais contribuiu foi linfonodomegalia, seguido de idade, sexo e perda ponderal. O peso desta última variável pode ser explicado pelo fato destes tumores frequentemente comprometerem a alimentação, devido ao acometimento da base da língua, importante região para transferência do alimento da cavidade oral para orofaringe. Além disso, acreditamos que a perda ponderal esteja relacionada a redução da massa muscular esquelética, redução da reserva energética do paciente e da capacidade do sistema imune em combater infecções. Tais consequências aumentam os riscos do tratamento oncológico e, muitas vezes, levam a repetidas internações hospitalares. Alguns estudos já sugeriram a perda ponderal antes

do tratamento como importante fator prognóstico(132). No estudo brasileiro de Baptistella *et al.*, 77% dos pacientes que perderam mais que cinco quilos morreram em cinco anos, contra 13% para aqueles que perderam no máximo cinco quilos ($p=0,0001$)(133).

Dentre todos os sítios, sem dúvida a construção de um modelo para hipofaringe foi o mais desafiador. Nossa população tinha apenas 167 casos de hipofaringe, com poucos sobreviventes, e sabíamos da dificuldade em construir, validar e testar um modelo com menos casos. Levando em conta os preceitos de Harrell *et al.*(66) seria possível a construção de um modelo composto apenas por três variáveis, uma vez que 35 pacientes sobreviveram ao quinto ano (10 eventos por variável). Os modelos construídos tiveram ASC excelente na subamostra de treinamento ($ASC=0,89$), mas falharam no teste ($ASC=0,45$). Isto demonstra a importância de sempre separar parte da amostra para teste ao construir um modelo, pois na maioria das vezes que uma modelo preditivo falha na validação externa, já teria falhado caso fosse submetido a uma validação interna. Essa discrepância também nos faz refletir acerca da quantidade de estudos tomados como verdade, que assumiram como válido o desempenho do modelo apenas em sua amostra de treinamento. Na análise do modelo CART encontramos que a perda ponderal foi a variável que melhor discriminou os grupos (vivo/morto) dentre todas as variáveis.

O segundo sítio com maior número de pacientes era o de tumores de laringe. É já estabelecida a diferença de comportamento entre tumores nas diferentes sub-regiões da laringe – supraglote, glote e subglote. Segundo Ding *et al*, carcinomas supraglóticos possuem características clínicas distintas quando comparado aos outros subsítios laríngeos, como expressão do HPV e frequência de metástases linfonodais, que impactam no prognóstico destes pacientes (134). Contudo, foi difícil utilizar esta informação. Muitos tumores eram avançados e acometiam mais de uma sub-região, ou seja, eram considerados transglóticos. Houve muita heterogeneidade entre os diferentes centros coletores de dados acerca do modo como esses tumores avançados eram alocados (ou não) a uma das sub-regiões da laringe. Por esse motivo, optamos por explorar fatores de risco em uma amostra de pacientes mais robusta, porém heterogênea. Já de início, observamos que os tumores de laringe diferiam dos tumores de outros sítios pelo menor volume clínico. Acreditamos que isso se explique pelo fato de os tumores da laringe comprometerem mais precocemente a permeabilidade da via aérea, o que motiva a busca

mais pronta por tratamento pelos pacientes. Ao construir um modelo de CART, identificamos um novo ponto de corte para volume de 3cm^3 , diferentemente de 25cm^3 para todos os sítios. Assim, criamos uma variável clínica binária exclusiva para tumores laríngeos com este ponto de corte. Alguns estudos já demonstraram a importância prognóstica do volume tumoral para tumores laríngeos(135). Miyabe *et al.* observaram que o volume tumoral clínico mensurado através de exame de imagem é superior ao estadiamento T do TNM para prever sobrevida global(136).

Ainda em relação aos tumores laríngeos, o grau de diferenciação foi o parâmetro que melhor contribuiu para o modelo de BRT. A correlação entre o grau de diferenciação em câncer de laringe com a sobrevida global já foi sugerida em estudos prévios, principalmente levando em consideração o fronte de invasão tumoral(137, 138). De acordo com a literatura, pacientes com tumores pouco diferenciados são mais propensos a desenvolver metástases regionais e à distância o que explicaria o pior prognóstico(139).

Por fim, no modelo de regressão logística para tumores laríngeos, as variáveis preditivas com significância estatística foram idade (OR=1,05; IC 95% 1,01-1,09), linfonodomegalia (OR=2,37; IC 95% 1,03-5,65) e volume clínico binário (com ponto de corte em 3cm^3 ; OR=4,66; IC 95% 2,15-10,55). Apesar do modelo final ter apresentando uma ASC satisfatória (ASC=0,71) e passado nos testes de calibração, entendemos que uma das limitações da nossa análise foi não termos disponível a informação do subsítio laríngeo.

A mineração de dados empregando os métodos baseados em árvores associado ao conhecimento clínico progresso de possíveis fatores preditivos permitiu a construção dos modelos. Comparações envolvendo o TNM foram realizadas através de curvas de sobrevida. Utilizando os modelos de regressão logística e redes neurais construídos, observamos uma maior separação entre as curvas determinadas pelos quartis Q1-4 de risco previsto pelos respectivos modelos, em comparação ao TNM. Estes achados foram observados em todos os modelos, exceto aqueles para hipofaringe, os quais não apresentaram curvas de sobrevidas satisfatórias, o mesmo encontrado quando utilizamos o sistema TNM para este sítio. Estes resultados mostraram que foi possível criar um modelo multivariado superior ao TNM em predição de mortalidade em cinco anos.

O sistema TNM perde bastante acurácia preditiva devido a formação de grandes grupos baseados apenas em características clínicas do tumor primário e presença de metástases regionais ou à distância, ou seja, perde-se a capacidade preditiva ao agrupar pacientes tão heterogêneos em um mesmo grupo de risco(9, 140). Nos modelos criados durante este estudo, os quatro grupos de risco (Q1, Q2, Q3, e Q4) foram definidos de acordo com a faixa de risco de mortalidade. A grande vantagem do método multivariado é a presença da covariância, permitindo que fatores façam parte, ou não, dos modelos, além de ajustar o peso de cada variável.

Sabe-se que a expressão do HPV em tumores de orofaringe está fortemente associada a melhor sobrevida e resposta terapêutica(24), por se tratarem de tumores de etiologia, epidemiologia e perfil biológico distintos quando comparados aos tumores causados pelo álcool e pelo cigarro. Por esses motivos, há algumas iniciativas em busca de modelos preditivos exclusivos para tumores com este perfil (58, 141, 142). Seu papel em outros subsítios da cabeça e pescoço, contudo, é questionável(26).

Dentre os 1811 pacientes da nossa amostra, 692 pacientes possuíam informação sobre o HPV e em 60 (8,7%) foi possível detectar a presença do vírus (subtipos de alto risco). Uma metanálise publicada em 2015(143), foi sugerido que 27,4% dos pacientes com CECp possuíam infecção de algum subtipo de HPV, sendo 8,0% com HPV 16. Em um estudo canadense publicado por Bratman *et al.*(144), foi identificado HPV de alto risco em 73 (14%) pacientes de uma amostra com 515 pacientes com CECp. A diferença entre a positividade do HPV na nossa população com a dos países desenvolvidos pode ser explicada pelo perfil dos pacientes. Nossa amostra é caracterizada por homens tabagistas na sexta década de vida, enquanto que nos países desenvolvidos o perfil é de pacientes mais jovens não tabagistas(145).

Para avaliar se a infecção pelo HPV tinha algum valor prognóstico em nossa população, criamos dois modelos de regressão para esta subamostra. No primeiro incluímos apenas as variáveis selecionadas anteriormente e no segundo incluímos a entrada do HPV no modelo. Utilizando a seleção passo a passo para selecionar variáveis, nenhum dos dois métodos (AIC ou *Max Validation R²*) incorporaram o HPV no modelo, o que demonstra que não foi uma variável importante para a predição pelo modelo. Para poder avaliar seu peso, forçamos a sua entrada na análise, porém não observamos melhora significativa da performance do modelo ao adicionar o HPV, com indícios de piora da

mesma após adição.

Dentre os 692 pacientes que possuíam informações da expressão do HPV, 119 pacientes possuíam como sítio primário a orofaringe e, destes, 14 foram positivos. Devido à pouca quantidade de pacientes, não foi possível subdividir a amostra em treinamento, validação e teste. Ainda assim, criamos dois modelos (com e sem o HPV), não sendo encontrado melhora da capacidade preditiva ao incluir o HPV. A ausência do incremento do HPV nos modelos preditivos construídos pode ser explicada pelo perfil predominantemente tabagista dos nossos pacientes. Ademais, apesar de esta ser uma das maiores séries de dados de HPV em CECp no Brasil, dada a baixa positividade, pode ainda sim ter sido insuficiente para nossos objetivos.

6.1 - Limitações

Mesmo contando com um banco de dados cujas informações foram coletadas prospectivamente por entrevistadores treinados, estudos retrospectivos contam com dificuldades adicionais. Conseguimos recuperar uma grande quantidade de dados perdidos e as análises de consistências evidenciaram alguns dados equivocados, sendo prontamente corrigidos. Por outro lado, uma parte das informações foram perdidas e não conseguimos recuperar, principalmente quando os pacientes foram incluídos por outro centro que não o HC-FMUSP, dada nossa dificuldade de acesso aos dados de outros hospitais. Levando em consideração que a regressão logística foi uma das ferramentas utilizadas na construção do modelo, a perda de dados de apenas uma variável já exclui o paciente do cálculo, consequentemente ocasionando uma redução do número de pacientes nas subamostras de treinamento, validação e teste.

Já está bem estabelecido o impacto da presença de comorbidade na sobrevida global em pacientes com CECp, sobretudo em idosos(7, 8, 146). Segundo Datema *et al*, a comorbidade impacta na sobrevida global dos pacientes portadores de CECp e a presença de uma comorbidade grave é equivalente a um estágio T4N2 do TNM(147). Devido à indisponibilidade desta informação no banco de dados GENCAPO, este parâmetro não foi utilizado como preditor. O grupo de pesquisa já incorporou essas informações aos questionários dos pacientes incluídos nas fases seguintes a este projeto (GENCAPO2).

Com o objetivo de otimizar os recursos, aproveitamos dados de pesquisas do HPV de estudos prévios realizados pelo GENCAPO. Com isto a variável relacionada ao HPV contou com dois métodos distintos: hibridização reversa (INNO-LIPA[®]) e reação de imunogenotipagem (BD Onclarity[®]). Apesar de boa concordância entre os métodos, existem algumas diferenças, sendo a principal o fato da primeira ser realizada de forma manual e a segunda, automatizada. O único centro em que tivemos acesso aos blocos foi o HC-FMUSP e nem todos os blocos foram recuperados. Além disso alguns dos blocos encontrados possuíam pouco tecido paraafinado ou áreas não tumorais, dificultando as análises. O fato de possuímos resultados de HPV apenas de uma parte da amostra reduziu nosso poder estatístico, com isto avaliamos o incremento do HPV na predição utilizando todos os sítios. O modelo para orofaringe ficou instável devido à pouca quantidade de casos, não sendo criado.

A presença de dados faltantes é comum em pesquisa clínica, sobretudo em estudos retrospectivos. Realizamos imputação dos dados de dimensões clínicas utilizando como parâmetro medidas obtidas no estudo de anatomia patológica. Este critério não é o ideal uma vez que as dimensões clínicas tendem a ser subdimensionadas quando comparadas com as dimensões tumorais da peça cirúrgica. Para minimizar esse ponto, criamos a equação (imputação multivariada) para corrigir esta diferença.

Dados relacionados à perda ponderal podem ter sido coletados de forma imprecisa, pois nem sempre havia balanças disponíveis e a informação era obtida subjetivamente ao questionar o paciente.

Alguns pacientes perdiam seguimento e as informações para localizá-los, como telefone e endereço, estavam desatualizadas, uma vez que se trata de um estudo envolvendo pacientes entrevistados a partir de 2001. Grande parte dos dados foi recuperada, mas ainda assim perdemos dados de seguimento de alguns pacientes.

O último paciente inserido neste estudo foi em 2011, portanto antes do lançamento da 8ª edição do TNM. Re-estadiamos estes pacientes utilizando dados de espessura tumoral para cavidade oral (parte deles imputados) e dados de estudo de anatomia patológica em relação a extravasamento extracapsular. Apenas parte dos pacientes de orofaringe possuía informação do HPV, não sendo possível o re-estadiamento de todos, contudo, o próprio TNM contempla a falta de informação do status TNM.

Todos os pacientes incluídos neste estudo foram atendidos e tratados no Sistema Único de Saúde. Deve-se levar em consideração as limitações deste perfil de paciente como o atraso no diagnóstico e restrições relacionados ao tratamento. Em um modelo baseado em aprendizado de máquina foi identificado que o seguro saúde foi um dos principais fatores preditivos de morte em pacientes com CECP nos EUA(148). Em um estudo publicado por Kwok *et al*, pacientes com Medicaid ou sem seguro de saúde possuíram um risco aumentado de morte (OR=1,5; IC95% 1,07-2,11) após o diagnóstico de CECP quando comparado com pacientes com seguros de saúde privado, mesmo ajustando para idade, sexo, raça, tabagismo, etilismo, nível socioeconômico, tratamento e estágio pelo TNM(149). Isso ressalta a importância de fatores sociais e não apenas anatômicos para a predição de prognóstico em câncer.

A única variável que em nenhum momento contribuiu na construção dos modelos foi metástases a distância. Sabemos que metástase a distância é um importante fator preditivo de mortalidade(150) e acreditamos que ela não tenha entrado nos modelos devido à ocorrência muito baixa na nossa população. Tal fato se deve provavelmente a falta de diagnóstico.

6.2 – Perspectivas futuras

Criar um modelo preditivo em oncologia capaz de prever a mortalidade em cinco anos com boa acurácia é extremamente útil, tanto para os médicos quanto para o paciente.

Nesta tese conseguimos criar modelos preditivos para cavidade oral, orofaringe e laringe com discriminação satisfatória, excelente calibração e superiores à última e penúltima edições do TNM quando comparamos as curvas de sobrevida.

Todos os modelos passaram pelo processo de validação e teste em amostras distintas, entretanto provenientes de uma mesma população. Como próximo passo, devemos aplicar estes modelos em outras populações visando validá-los externamente.

Uma vez validado em outras populações, criaremos um software de manuseio simples com o objetivo de disponibilizar esta ferramenta na prática clínica diária.

O desenvolvimento de ferramentas preditivas é um processo sem fim. Novos estudos apresentarão variáveis desconhecidas até o momento, levando a necessidade de atualização dos modelos aqui apresentados.

CONCLUSÃO

7.0 - CONCLUSÃO

Identificamos parâmetros não contemplados pelo TNM significativos na predição de risco de morte em cinco anos para pacientes com carcinoma epidermoide de cabeça e pescoço.

Criamos modelos de predição mais precisos que o TNM para cavidade oral, orofaringe e laringe, além de um modelo independente do sítio tumoral primário.

Avaliamos o incremento do status HPV em modelos de predição de risco de morte em cinco anos independente do sítio tumoral primário e em pacientes com tumores de orofaringe e não encontramos melhora na performance dos modelos.

ANEXOS

8.0 - ANEXOS

ANEXO A

Tabela A.1 – Distribuição dos pacientes por centro.

Centro	População GENCAPO n(%)	Desfecho faltante n(%)	Incluído nas análises n(%)
Instituto Arnaldo Vieira de Carvalho	365 (17,4)	33 (11,7)	332 (18,3)
Hospital Heliópolis	735 (35,1)	82 (29,2)	653 (36,1)
Hospital das Clínicas da Faculdade de Medicina da Universidade São Paulo	624 (29,8)	13 (4,6)	611 (33,7)
Universidade do Vale do Paraíba	18 (0,9)	16 (5,7)	2 (0,1)
Faculdade de Medicina da Universidade de São Paulo de Ribeirão Preto	350 (16,7)	137 (48,8)	213 (11,8)
Total	2092 (100)	281 (100)	1811 (100)

Tabela A.2 – Distribuição sociodemográfica.

Característica		População GENCAPO n (%)	Desfecho faltante n (%)	Incluído nas análises n (%)
Gênero	Masculino	1807 (86,6)	244 (87,1)	1563 (86,5)
	Feminino	280 (13,4)	36 (12,9)	244 (13,5)
Raça	Branca	1468 (71,2)	194 (71,3)	1274 (71,1)
	Mulata	386 (18,7)	47 (17,3)	339 (18,9)
	Negra	183 (8,9)	29 (10,7)	154 (8,6)
	Oriental	18 (0,9)	0 (0)	18 (0,1)
	Indígena	8 (0,4)	2 (0,7)	6 (0,003)
Escolaridade	Analfabeto	245 (12,0)	25 (10,0)	220 (12,3)
	Alfabetizado	73 (3,6)	7 (2,8)	66 (3,7)
	Primário Incompleto	591 (28,9)	74 (29,5)	517 (28,9)
	Primário Completo	686 (33,6)	109 (43,4)	577 (32,2)
	Secundário Incompleto	125 (6,1)	5 (2,0)	120 (6,7)
	Secundário Completo	232 (11,4)	20 (8,0)	212 (11,8)
	Técnico	35 (1,8)	2 (0,8)	33 (1,8)
	Universitário	55 (2,7)	9 (3,6)	46 (2,6)
Tabagismo	Nunca fumou	129 (6,3)	10 (3,9)	119 (6,7)
	Tabagista	1441 (70,1)	180 (69,8)	1261 (70,9)
	Extabagista	465 (22,9)	68 (26,4)	397 (22,3)
Etilismo	Nunca	263 (13,0)	24 (9,4)	239 (13,4)
	Etilista	1030 (50,8)	130 (51,2)	900 (50,8)
	Ex-etilista	733 (36,2)	100 (39,3)	633 (35,7)

ANEXO A

Tabela A.3 – Distribuição dos sintomas.

Característica		População GENCAPO n (%)	Desfecho faltante n (%)	Incluído nas análises n (%)
Perda Ponderal	Sim	1199 (59,8)	126 (52,3)	1073 (60,8)
	Não	806 (40,2)	115 (47,8)	691 (39,2)
Disfonia	Sim	799 (39,4)	115 (45,0)	684 (38,6)
	Não	1228 (60,6)	141 (55,1)	1087 (61,4)
Odinofagia	Sim	677 (33,7)	90 (36,3)	587 (33,2)
	Não	1337 (66,4)	158 (63,7)	1179 (66,8)

Tabela A.4 – Distribuição de acordo com idade (anos).

Quartis		População GENCAPO n (%)	Desfecho faltante n (%)	Incluído nas análises n (%)
100.0%	Máximo	93	87	93
99.5%		86	86,18	86
97.5%		79	80	79
90.0%		72	71	72
75.0%	quartil	65	65	65
50.0%	mediana	57	57	57
25.0%	quartil	50	50	50
10.0%		45	44	45
2.5%		39	40	39
0.5%		30,46	30,82	30,06
0.0%	mínimo	21	30	21

	População GENCAPO n (%)	Desfecho faltante n (%)	Incluído nas análises n (%)
Média	57,67	57,69	57,67
Desvio Padrão	10,37	10,47	10,36
Médio do Erro Padrão	0,23	0,62	0,24
Superior 95% da média	58,11	58,92	58,15
Inferior 95% da média	57,22	56,46	57,19
N	2091	281	1811

ANEXO A

Tabela A.5 – Distribuição de acordo com a perda ponderal (%) - apenas pacientes incluídos nas análises.

100.0%	máximo	40,54
99.5%		34,55
97.5%		25,71
90.0%		16,7
75.0%	quartil	10
50.0%	mediana	0,68
25.0%	quartil	0,00
10.0%		0,00
2.5%		0,00
0.5%		0,00
0.0%	mínimo	0,000
Média		5,77
Desvio Padrão		7,69
Erro de Desvio Padrão		0,19
Superior 95% da média		6,14
Inferior 95% da média		5,40
N		1692

ANEXO A

Tabela A.6 – Distribuição de acordo com a quantidade de tempo que houve a perda ponderal (em meses) – apenas pacientes incluídos nas análises.

100.0%	máximo	108
99.5%		42,39
97.5%		24
90.0%		12
75.0%	quartil	6
50.0%	mediana	4
25.0%	quartil	2
10.0%		2
2.5%		1
0.5%		1
0.0%	mínimo	1
Média		5,39
Desvio Padrão		6,14
Erro de Desvio Padrão		0,19
Superior 95% da média		5,77
Inferior 95% da média		5,02
N		1057

ANEXO A

Tabela A.7- Distribuição de acordo com o índice de massa corpórea (IMC) - apenas pacientes incluídos nas análises.

100.0%	máximo	42,45
99.5%		37,63
97.5%		33,06
90.0%		29,07
75.0%	quartil	26,3
50.0%	mediana	23,61
25.0%	quartil	21,3
10.0%		19,49
2.5%		17,10
0.5%		15,23
0.0%	mínimo	12,91
Média	24,00	
Desvio Padrão	3,97	
Erro de Desvio Padrão	0,098	
Superior 95% da média	24,20	
Inferior 95% da média	23,81	
N	1639	

ANEXO B

Tabela B.1 – Dados obtidos após construção do modelo baseado em redes neurais para tumores todos os sítios.

	Treinamento	Validação	Teste
R ² Geral	0,26	0,17	0,24
R ² Entropia	0,17	0,10	0,15
-LogVerossimilhança	484,17	257,10	267,27
n	905	435	471

Tabela B.2 – Regressão logística para tumores de todos os sítios.

Medidas	Treinamento	Validação	Teste
Entropia R ²	0,14	0,12	0,16
R ² Generalizado	0,22	0,20	0,26
Média -Log p	0,56	0,59	0,57
N	722	348	383

Tabela B.3 – Dados obtidos após construção do modelo baseado em redes neurais para tumores de cavidade oral.

	Treinamento	Validação	Teste
R ² Geral	0,21	0,07	0,19
R ² Entropia	0,13	0,04	0,11
-LogVerossimilhança	210,41	111,12	122,30
N	369	177	206

Tabela B.4 – Regressão logística para tumores de cavidade oral.

Medidas	Treinamento	Validação	Teste
Entropia R ²	0,12	0,04	0,15
R ² Generalizado	0,19	0,07	0,25
Média -Log p	0,58	0,64	0,57
N	318	158	179

ANEXO B

Tabela B.5 – Dados obtidos após construção do modelo baseado em redes neurais para tumores de orofaringe.

	Treinamento	Validação	Teste
R ² Geral	0,2893	0,1778	-0,1275
R ² Entropia	0,1896	0,1087	-0,0687
-LogVerossimilhança	75,8815	38,8364	52,1769
n	169	70	76

Tabela B.6 – Regressão logística para tumores de orofaringe.

Medidas	Treinamento	Validação	Teste
Entropia R ²	0,26	-0,03	-0,01
R ² Generalizado	0,38	-0,05	-0,02
Média -Log p	0,46	0,54	0,63
N	144	111	118

Tabela B.7 – Dados obtidos após construção do modelo baseado em redes neurais hipofaringe.

	Treinamento	Validação	Teste
R ² Geral	-0,46	0,24	0,26
R ² Entropia	-0,25	0,15	0,14
-LogVerossimilhança	24,76	13,03	8,53
n	39	24	22

Tabela B.8 – Regressão logística para tumores de hipofaringe.

Medidas	Treinamento	Validação	Teste
Entropia R ²	0,36	-1,29	-0,56
R ² Generalizado	0,46	-5,09	-1,25
Média -Log p	0,27	1,33	0,85
N	47	56	47

ANEXO B

Tabela B.9 – dados obtidos após construção do modelo baseado em redes neurais para tumores de laringe.

	Treinamento	Validação	Teste
R ² Geral	0,30	0,21	0,23
R ² Entropia	0,18	0,13	0,14
-LogVerossimilhança	106,50	53,95	59,84
n	189	90	100

Tabela B.10 – Regressão logística para tumores de laringe.

Medidas	Treinamento	Validação	Teste
Entropia R ²	0,17	-0,02	0,05
R ² Generalizado	0,28	-0,03	0,10
Média -Log p	0,30	0,41	0,36
N	158	143	122

REFERÊNCIAS

9.0 – REFERÊNCIAS:

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin.* 2018;68(6):394-424.
2. INCA - Instituto Nacional de Câncer - Estimativa 2018.
3. Amin MB, Greene FL, Edge SB, Compton CC, Gershenwald JE, Brookland RK, et al. The Eighth Edition AJCC Cancer Staging Manual: Continuing to build a bridge from a population-based to a more "personalized" approach to cancer staging. *CA Cancer J Clin.* 2017;67(2):93-9.
4. Lydiatt WM, Patel SG, O'Sullivan B, Brandwein MS, Ridge JA, Migliacci JC, et al. Head and Neck cancers-major changes in the American Joint Committee on cancer eighth edition cancer staging manual. *CA Cancer J Clin.* 2017;67(2):122-37.
5. Wolfensberger M. Using Cox's proportional hazards model for prognostication in carcinoma of the upper aero-digestive tract. *Acta Otolaryngol.* 1992;112(2):376-82.
6. van der Schroeff MP, Derks W, Hordijk GJ, de Leeuw RJ. The effect of age on survival and quality of life in elderly head and neck cancer patients: a long-term prospective study. *Eur Arch Otorhinolaryngol.* 2007;264(4):415-22.
7. Hall SF, Groome PA, Rothwell D. The impact of comorbidity on the survival of patients with squamous cell carcinoma of the head and neck. *Head & Neck.* 2000;22(4):317-22.
8. Reid BC, Alberg AJ, Klassen AC, Samet JM, Rozier RG, Garcia I, et al. Comorbidity and survival of elderly head and neck carcinoma patients. *Cancer.* 2001;92(8):2109-16.
9. Baatenburg de Jong RJ, Hermans J, Molenaar J, Briaire JJ, le Cessie S. Prediction of survival in patients with head and neck cancer. *Head & Neck.* 2001;23(9):718-24.
10. Burke HB, Goodman PH, Rosen DB, Henson DE, Weinstein JN, Harrell FE, et al. Artificial neural networks improve the accuracy of cancer survival prediction. *Cancer.* 1997;79(4):857-62.
11. López RVM, Levi JE, Eluf-Neto J, Koifman RJ, Koifman S, Curado MP, et al. Human papillomavirus (HPV) 16 and the prognosis of head and neck cancer in a geographical region with a low prevalence of HPV infection. *Cancer Causes Control.* 2014;25(4):461-71.
12. Tajara EH, Villa LL, Carvalho MBd. Research interests: behind the biological sample. *BMJ.* 2014;349:g5231.
13. Marur S, Forastiere AA. Head and Neck Cancer: Changing Epidemiology, Diagnosis, and Treatment. *Mayo Clinic Proceedings.* 2008;83(4):489-501.
14. Sturgis EM, Cinciripini PM. Trends in head and neck cancer incidence in relation to smoking prevalence. *Cancer.* 2007;110(7):1429-35.
15. Chaturvedi AK, Engels EA, Pfeiffer RM, Hernandez BY, Xiao W, Kim E, et al. Human Papillomavirus and Rising Oropharyngeal Cancer Incidence in the United States. *J Clin Oncol.* 2011;29(32):4294-301.
16. D'Souza G, Anantharaman D, Gheit T, Abedi-Ardekani B, Beachler DC, Conway DI, et al. Effect of HPV on head and neck cancer patient survival, by region and tumor site: A comparison of 1362 cases across three continents. *Oral Oncol.* 2016;62:20-7.
17. Ferlay J, Soerjomataram I, Dikshit R, Eser S, Mathers C, Rebelo M, et al. Cancer incidence and mortality worldwide: Sources, methods and major patterns in GLOBOCAN 2012. *Int J Cancer.* 2015;136(5):E359-E86.
18. Torre LA, Bray F, Siegel RL, Ferlay J, Lortet-Tieulent J, Jemal A. Global cancer statistics, 2012. *CA Cancer J Clin.* 2015;65(2):87-108.

19. Sobin LH. Tnm. Cancer. 2001;91(S8):1589-92.
20. Huang SH, O'Sullivan B. Overview of the 8th Edition TNM Classification for Head and Neck Cancer. *Curr Treat Options Oncol*. 2017;18(7):40.
21. Vickers AJ. Prediction models in cancer care. *CA Cancer J Clin*. 2011;61(5):315-26.
22. Dahlstrom KR, Calzada G, Hanby JD, Garden AS, Glisson BS, Li G, et al. An evolution in demographics, treatment, and outcomes of oropharyngeal cancer at a major cancer center: a staging system in need of repair. *Cancer*. 2013;119(1):81-9.
23. Spector ME, Gallagher KK, Bellile E, Chinn SB, Ibrahim M, Byrd S, et al. Patterns of nodal metastasis and prognosis in human papillomavirus-positive oropharyngeal squamous cell carcinoma. *Head Neck*. 2014;36(9):1233-40.
24. Huang SH, Xu W, Waldron J, Siu L, Shen X, Tong L, et al. Refining American Joint Committee on Cancer/Union for International Cancer Control TNM stage and prognostic groups for human papillomavirus-related oropharyngeal carcinomas. *J Clin Oncol*. 2015;33(8):836-45.
25. O'Sullivan B, Huang SH, Su J, Garden AS, Sturgis EM, Dahlstrom K, et al. Development and validation of a staging system for HPV-related oropharyngeal cancer by the International Collaboration on Oropharyngeal cancer Network for Staging (ICON-S): a multicentre cohort study. *Lancet Oncol*. 2016;17(4):440-51.
26. Porceddu SV, Milne R, Brown E, Bernard A, Rahbari R, Cartmill B, et al. Validation of the ICON-S staging for HPV-associated oropharyngeal carcinoma using a pre-defined treatment policy. *Oral Oncol*. 2017;66:81-6.
27. Würdemann N, Wagner S, Sharma SJ, Prigge ES, Reuschenbach M, Gattenlöhner S, et al. Prognostic Impact of AJCC/UICC 8th Edition New Staging Rules in Oropharyngeal Squamous Cell Carcinoma. *Front Oncol*. 2017;7.
28. Mizumachi T, Homma A, Sakashita T, Kano S, Hatakeyama H, Fukuda S. Confirmation of the eighth edition of the AJCC/UICC TNM staging system for HPV-mediated oropharyngeal cancer in Japan. *Int J Clin Oncol*. 2017;22(4):682-9.
29. van Gysen K, Stevens M, Guo L, Jayamanne D, Veivers D, Wignall A, et al. Validation of the 8(th) edition UICC/AJCC TNM staging system for HPV associated oropharyngeal cancer patients managed with contemporary chemo-radiotherapy. *BMC Cancer*. 2019;19(1):674.
30. Choi SH, Terrell JE, Fowler KE, McLean SA, Ghanem T, Wolf GT, et al. Socioeconomic and Other Demographic Disparities Predicting Survival among Head and Neck Cancer Patients. *PLOS ONE*. 2016;11(3):e0149886.
31. Wynder EL, Bross IJ, Feldman RM. A study of the etiological factors in cancer of the mouth. *Cancer*. 1957;10(6):1300-23.
32. Hammond EC, Horn D. Smoking and death rates: report on forty-four months of follow-up of 187,783 men. 2. Death rates by cause. *J Am Med Assoc*. 1958;166(11):1294-308.
33. Humans IWGotEoCRt. Tobacco Smoke and Involuntary Smoking: International Agency for Research on Cancer; 2004 2004.
34. Pytynia KB, Grant JR, Etzel CJ, Roberts DB, Wei Q, Sturgis EM. Matched-pair analysis of survival of never smokers and ever smokers with squamous cell carcinoma of the head and neck. *J Clin Oncol*. 2004;22(19):3981-8.
35. Farshadpour F, Kranenborg H, Calkoen EVB, Hordijk GJ, Koole R, Slootweg PJ, et al. Survival analysis of head and neck squamous cell carcinoma: influence of smoking and drinking. *Head & Neck*. 2011;33(6):817-23.
36. Ragin CC, Langevin SM, Marzouk M, Grandis JR, Taioli E. Determinants of head

- and neck cancer survival according to race. *Head & neck*. 2011;33(8):1092-8.
37. Molina MA, Cheung MC, Perez EA, Byrne MM, Franceschi D, Moffat FL, et al. African American and poor patients have a dramatically worse prognosis for head and neck cancer. *Cancer*. 2008;113(10):2797-806.
 38. Morse DE, Kerr AR. Disparities in oral and pharyngeal cancer incidence, mortality and survival among black and white Americans. *J Am Dent Assoc*. 2006;137(2):203-12.
 39. Callegari-Jacques SM, Grattapaglia D, Salzano FM, Salamoni SP, Crossetti SG, Ferreira ME, et al. Historical genetics: Spatiotemporal analysis of the formation of the Brazilian population. *Am J Hum Biol*. 2003;15(6):824-34.
 40. Pimenta JR, Zuccherato LW, Debes AA, Maselli L, Soares RP, Moura-Neto RS, et al. Color and genomic ancestry in Brazilians: a study with forensic microsatellites. *Hum Hered*. 2006;62(4):190-5.
 41. Camilon PR, Stokes WA, Nguyen SA, Lentsch EJ. The prognostic significance of age in oropharyngeal squamous cell carcinoma. *Oral Oncol*. 2014;50(5):431-6.
 42. Udeabor SE, Rana M, Wegener G, Gellrich N-C, Eckardt AM. Squamous cell carcinoma of the oral cavity and the oropharynx in patients less than 40 years of age: a 20-year analysis. *Head & Neck Oncology*. 2012;4:28.
 43. Claus EB, Risch N, Thompson WD. Autosomal dominant inheritance of early-onset breast cancer. Implications for risk prediction. *Cancer*. 1994;73(3):643-51.
 44. Mazzola E, Chipman J, Cheng S-C, Parmigiani G. Recent BRCAPRO upgrades significantly improve calibration. *Cancer Epidemiol Biomarkers Prev*. 2014;23(8):1689-95.
 45. Gail MH, Brinton LA, Byar DP, Corle DK, Green SB, Schairer C, et al. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *Journal of the National Cancer Institute*. 1989;81(24):1879-86.
 46. Lee AJ, Cunningham AP, Kuchenbaecker KB, Mavaddat N, Easton DF, Antoniou AC. BOADICEA breast cancer risk prediction model: updates to cancer incidences, tumour pathology and web interface. *British Journal of Cancer*. 2014;110(2):535-45.
 47. Thompson IM, Ankerst DP, Chi C, Goodman PJ, Tangen CM, Lucia MS, et al. Assessing Prostate Cancer Risk: Results from the Prostate Cancer Prevention Trial. *J Natl Cancer Inst*. 2006;98(8):529-34.
 48. Roobol MJ, van Vugt HA, Loeb S, Zhu X, Bul M, Bangma CH, et al. Prediction of Prostate Cancer Risk: The Role of Prostate Volume and Digital Rectal Examination in the ERSPC Risk Calculators. *European Urology*. 2012;61(3):577-83.
 49. D'Amico AV. Biochemical Outcome After Radical Prostatectomy, External Beam Radiation Therapy, or Interstitial Radiation Therapy for Clinically Localized Prostate Cancer. *JAMA*. 1998;280(11):969.
 50. Cooperberg MR, Pasta DJ, Elkin EP, Litwin MS, Latini DM, DuChane J, et al. The UCSF Cancer of the Prostate Risk Assessment (CAPRA) Score: a straightforward and reliable preoperative predictor of disease recurrence after radical prostatectomy. *J Urol*. 2005;173(6):1938-42.
 51. Stephenson AJ, Scardino PT, Eastham JA, Bianco FJ, Dotan ZA, Fearn PA, et al. Preoperative Nomogram Predicting the 10-Year Probability of Prostate Cancer Recurrence After Radical Prostatectomy. *J Natl Cancer Inst*. 2006;98(10):715-7.
 52. Eggener SE, Scardino PT, Walsh PC, Han M, Partin AW, Trock BJ, et al. PREDICTING FIFTEEN-YEAR CANCER-SPECIFIC MORTALITY BASED ON THE PATHOLOGICAL FEATURES OF PROSTATE CANCER. *J Urol*. 2011;185(3):869-75.

53. Mamoshina P, Vieira A, Putin E, Zhavoronkov A. Applications of Deep Learning in Biomedicine. *Mol Pharm.* 2016;13(5):1445-54.
54. Greene CS, Troyanskaya OG. Chapter 2: Data-driven view of disease biology. *PLoS Comput Biol.* 2012;8(12):e1002816.
55. Libbrecht MW, Noble WS. Machine learning applications in genetics and genomics. *Nat Rev Genet.* 2015;16(6):321-32.
56. Montazeri M, Montazeri M, Montazeri M, Beigzadeh A. Machine learning models in breast cancer survival prediction. *Technol Health Care.* 2016;24(1):31-42.
57. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature.* 2015;521(7553):436-44.
58. Qian X, Nguyen DT, Dong Y, Sinikovic B, Kaufmann AM, Myers JN, et al. Prognostic Score Predicts Survival in HPV-Negative Head and Neck Squamous Cell Cancer Patients. *Int J Biol Sci.* 2019;15(7):1336-44.
59. Fakhry C, Zhang Q, Nguyen-Tan PF, Rosenthal DI, Weber RS, Lambert L, et al. Development and Validation of Nomograms Predictive of Overall and Progression-Free Survival in Patients With Oropharyngeal Cancer. *J Clin Oncol.* 2017;35(36):4057-65.
60. Emerick KS, Leavitt ER, Michaelson JS, Diephuis B, Clark JR, Deschler DG. Initial clinical findings of a mathematical model to predict survival of head and neck cancer. *Otolaryngol Head Neck Surg.* 2013;149(4):572-8.
61. Ribeiro IP, Caramelo F, Esteves L, Menoita J, Marques F, Barroso L, et al. Genomic predictive model for recurrence and metastasis development in head and neck squamous cell carcinoma patients. *Sci Rep.* 2017;7(1):13897.
62. Moons KGM, Royston P, Vergouwe Y, Grobbee DE, Altman DG. Prognosis and prognostic research: what, why, and how? *BMJ.* 2009;338:b375.
63. Shickel B, Tighe PJ, Bihorac A, Rashidi P. Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis. *IEEE J Biomed Health Inform.* 2018;22(5):1589-604.
64. Prather JC, Lobach DF, Goodwin LK, Hales JW, Hage ML, Hammond WE. Medical data mining: knowledge discovery in a clinical data warehouse. *Proc AMIA Annu Fall Symp.* 1997:101-5.
65. Peduzzi P, Concato J, Feinstein AR, Holford TR. Importance of events per independent variable in proportional hazards regression analysis II. Accuracy and precision of regression estimates. *Journal of Clinical Epidemiology.* 1995;48(12):1503-10.
66. Harrell FE, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med.* 1996;15(4):361-87.
67. Altman DG. Categorising continuous variables. *British Journal of Cancer.* 1991;64(5):975.
68. Altman DG, Royston P. The cost of dichotomising continuous variables. *BMJ : British Medical Journal.* 2006;332(7549):1080.
69. Ringner M. What is principal component analysis? *Nat Biotechnol.* 2008;26(3):303-4.
70. Donders ART, van der Heijden GJMG, Stijnen T, Moons KGM. Review: A gentle introduction to imputation of missing values. *Journal of Clinical Epidemiology.* 2006;59(10):1087-91.
71. Moons KGM, Kengne AP, Woodward M, Royston P, Vergouwe Y, Altman DG, et al. Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio)marker. *Heart.* 2012;98(9):683-90.
72. Dong Y, Peng C-YJ. Principled missing data methods for researchers.

SpringerPlus. 2013;2(1):1-17.

73. Schafer JL. Multiple imputation: a primer. *Stat Methods Med Res.* 1999;8(1):3-15.
74. Bennett DA. How can I deal with missing data in my study? *Aust N Z J Public Health.* 2001;25(5):464-9.
75. Tabachnick BG, Fidell LS. *Using Multivariate Statistics.* 6th Edition ed2012.
76. Greenland S, Finkle WD. A critical look at methods for handling missing covariates in epidemiologic regression analyses. *Am J Epidemiol.* 1995;142(12):1255-64.
77. Little RJA. Regression With Missing X's: A Review. *Journal of the American Statistical Association.* 1992;87(420):1227-37.
78. Bertrand PV. *Theoretical epidemiology: Principles of Occurrence Research in Medicine.* by Olli S. Miettinen. John Wiley & Sons, New York, 1985. xxi + 359pp. ISBN 0-471-86769-1. *Can J Statistics.* 1988;16(3):317-8.
79. Perneger TV. What's wrong with Bonferroni adjustments. *Bmj.* 1998;316(7139):1236-8.
80. Dmitrienko A, D'Agostino R, Sr. Traditional multiplicity adjustment methods in clinical trials. *Stat Med.* 2013;32(29):5172-218.
81. Streiner DL, Norman GR. Correction for multiple testing: is there a resolution? *Chest.* 2011;140(1):16-8.
82. Smith AE, Mason AK. Cost Estimation Predictive Modeling: Regression Versus Neural Network. *The Engineering Economist.* 1997;42(2):137-61.
83. Steyerberg EW, Eijkemans MJC, Frank E. Harrell J, Habbema JDF. Prognostic Modeling with Logistic Regression Analysis: In Search of a Sensible Strategy in Small Data Sets. *Medical Decision Making.* 2001;21(1):45-56.
84. Setyawati BR, Sahirman S, Creese R. Neural networks for cost estimation. *AACE International Transactions of the Annual Meeting.* 2002:EST131-EST9.
85. Press SJ, Wilson S. Choosing between Logistic Regression and Discriminant Analysis. *Journal of the American Statistical Association.* 1978;73(364):699-705.
86. Weigend ASBWL. Bayesian Back-propagation. *Complex Systems.* 1991;5:603-43.
87. Menard S. *Applied Logistic Regression Analysis.* 2nd edition ed. Thousand Oaks, Calif: SAGE Publications, Inc; 2001 2001/10/09/. 128 p.
88. De'ath G. Boosted Trees for Ecological Modeling and Prediction. *Ecology.* 2007;88(1):243-51.
89. Austin PC, Lee DS, Steyerberg EW, Tu JV. Regression trees for predicting mortality in patients with cardiovascular disease: what improvement is achieved by using ensemble-based methods? *Biom J.* 2012;54(5):657-73.
90. Cox DR. Regression Models and Life-Tables. *Journal of the Royal Statistical Society Series B (Methodological).* 1972;34(2):187-220.
91. Ranganathan P, Pramesh CS, Aggarwal R. Common pitfalls in statistical analysis: Logistic regression. *Perspect Clin Res.* 2017;8(3):148-51.
92. Sperandei S. Understanding logistic regression analysis. *Biochem Med (Zagreb).* 2014;24(1):12-8.
93. Steyerberg EW, Eijkemans MJC, Frank E. Harrell, Jr., Habbema JDF. Prognostic Modeling with Logistic Regression Analysis: In Search of a Sensible Strategy in Small Data Sets. *Med Decis Making.* 2001;21(1):45-56.
94. Steyerberg E. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating.* New York: Springer-Verlag; 2009 2009.

95. Akaike H. Fitting autoregressive models for prediction. 1969.
96. Zhang Z. Variable selection with stepwise and best subset approaches. *Ann Transl Med.* 2016;4(7).
97. Cross SS, Harrison RF, Kennedy RL. Introduction to neural networks. *Lancet.* 1995;346(8982):1075-9.
98. Lippmann R. An introduction to computing with neural nets. *IEEE ASSP Magazine.* 1987;4(2):4-22.
99. Larentiis MD, Ravdin PM. Survival analysis of censored data: Neural network analysis detection of complex interactions between variables. *Breast Cancer Research and Treatment.* 2019;32(1):113-8.
100. Jones AS, Taktak AG, Helliwell TR, Fenton JE, Birchall MA, Husband DJ, et al. An artificial neural network improves prediction of observed survival in patients with laryngeal squamous carcinoma. *Eur Arch Otorhinolaryngol.* 2006;263(6):541-7.
101. Breiman L, Friedman J, Stone CJ, Olshen RA. Classification and Regression Trees. 1 edition ed. Boca Raton: Chapman and Hall/CRC; 1984. 368 p.
102. Chipman HA, George EI, McCulloch RE. Bayesian CART Model Search. *Journal of the American Statistical Association.* 1998;93(443):935-48.
103. Manning AM, Casper KA, Peter KS, Wilson KM, Mark JR, Collar RM. Can Predictive Modeling Identify Head and Neck Oncology Patients at Risk for Readmission? *Otolaryngol Head Neck Surg.* 2018;159(4):669-74.
104. van der Schroeff MP, Steyerberg EW, Wieringa MH, Langeveld TPM, Molenaar J, Baatenburg de Jong RJ. Prognosis: a variable parameter: dynamic prognostic modeling in head and neck squamous cell carcinoma. *Head & Neck.* 2012;34(1):34-41.
105. James G, Witten D, Hastie T, Tibshirani R. An Introduction to Statistical Learning 2013.
106. Datema FR, Ferrier MB, Vergouwe Y, Moya A, Molenaar J, Piccirillo JF, et al. Update and external validation of a head and neck cancer prognostic model. *Head Neck.* 2013;35(9):1232-7.
107. Hawkins DM. The problem of overfitting. *J Chem Inf Comput Sci.* 2004;44(1):1-12.
108. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology.* 2010;21(1):128-38.
109. Hosmer DW, Hosmer T, Le Cessie S, Lemeshow S. A Comparison of Goodness-of-Fit Tests for the Logistic Regression Model. *Statistics in Medicine.* 1997;16(9):965-80.
110. Hosmer DW, Lemeshow S. Applied Logistic Regression. 2nd edition ed. New York: Wiley-Interscience Publication; 2000 2000/09/15/. 392 p.
111. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Making.* 2006;26(6):565-74.
112. Draper NR, Smith H. Applied Regression Analysis. 3rd Edition ed 1998. 736 p.
113. Anantharaman D, Abedi-Ardekani B, Beachler DC, Gheit T, Olshan AF, Wisniewski K, et al. Geographic heterogeneity in the prevalence of human papillomavirus in head and neck cancer. *Int J Cancer.* 2017;140(9):1968-75.
114. Ejegod DM, Junge J, Franzmann M, Kirschner B, Bottari F, Sideri M, et al. Clinical and analytical performance of the BD Onclarity™ HPV assay for detection of CIN2+ lesions on SurePath samples. *Papillomavirus Research.* 2016;2:31-7.
115. Nogueira Dias Genta ML, Martins TR, Mendoza Lopez RV, Sadalla JC, de Carvalho JPM, Baracat EC, et al. Multiple HPV genotype infection impact on invasive

- cervical cancer presentation and survival. *PLOS ONE*. 2017;12(8):e0182854.
116. Castro FA, Koshiol J, Quint W, Wheeler CM, Gillison ML, Vaughan LM, et al. Detection of HPV DNA in paraffin-embedded cervical samples: a comparison of four genotyping methods. *BMC Infectious Diseases*. 2015;15.
 117. Walline HM, Komarek C, McHugh JB, Byrd SA, Spector ME, Hauff SJ, et al. High-risk human papillomavirus detection in oropharyngeal, nasopharyngeal, and, oral cavity cancers: Comparison of multiple methods. *JAMA Otolaryngol Head Neck Surg*. 2013;139(12):1320-7.
 118. Kreimer AR, Clifford GM, Boyle P, Franceschi S. Human papillomavirus types in head and neck squamous cell carcinomas worldwide: a systematic review. *Cancer Epidemiol Biomarkers Prev*. 2005;14(2):467-75.
 119. Monga SP, Wadleigh R, Sharma A, Adib H, Strader D, Singh G, et al. Intratumoral therapy of cisplatin/epinephrine injectable gel for palliation in patients with obstructive esophageal cancer. *Am J Clin Oncol*. 2000;23(4):386-92.
 120. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*. 1988;44(3):837-45.
 121. Groenwold RH, Donders AR, Roes KC, Harrell FE, Jr., Moons KG. Dealing with missing outcome data in randomized trials and observational studies. *Am J Epidemiol*. 2012;175(3):210-7.
 122. Tierney NJ, Harden FA, Harden MJ, Mengersen KL. Using decision trees to understand structure in missing data. *BMJ Open*. 2015;5(6):e007450.
 123. Xie K, Chen J, Zou J, Chen L, Gong L. Tumor volumes predict prognosis in head and neck cancer: A meta-analysis. *Translational Cancer Research*. 2017;6(4):687-97.
 124. DeJaco D, Steinbichler T, Schartinger VH, Fischer N, Anegg M, Dudas J, et al. Prognostic value of tumor volume in patients with head and neck squamous cell carcinoma treated with primary surgery. *Head Neck*. 2018;40(4):728-39.
 125. Beasley NJ, Prevo R, Banerji S, Leek RD, Moore J, van Trappen P, et al. Intratumoral lymphangiogenesis and lymph node metastasis in head and neck cancer. *Cancer Res*. 2002;62(5):1315-20.
 126. Cho JK, Hyun SH, Choi N, Kim MJ, Padera TP, Choi JY, et al. Significance of lymph node metastasis in cancer dissemination of head and neck cancer. *Transl Oncol*. 2015;8(2):119-25.
 127. Faye-Lund H, Abdelnoor M. Prognostic factors of survival in a cohort of head and neck cancer patients in Oslo. *Eur J Cancer B Oral Oncol*. 1996;32b(2):83-90.
 128. Roberts JC, Li G, Reitzel LR, Wei Q, Sturgis EM. No evidence of sex-related survival disparities among head and neck cancer patients receiving similar multidisciplinary care: a matched-pair analysis. *Clin Cancer Res*. 2010;16(20):5019-27.
 129. Medow MA, Weed HG, Schuller DE. Simple predictors of survival in head and neck squamous cell carcinoma. *Arch Otolaryngol Head Neck Surg*. 2002;128(11):1282-6.
 130. Huang SH, Hwang D, Lockwood G, Goldstein DP, O'Sullivan B. Predictive value of tumor thickness for cervical lymph-node involvement in squamous cell carcinoma of the oral cavity: a meta-analysis of reported studies. *Cancer*. 2009;115(7):1489-97.
 131. Ebrahimi A, Gil Z, Amit M, Yen T-C, Liao C-T, Chaturvedi P, et al. Primary Tumor Staging for Oral Cancer and a Proposed Modification Incorporating Depth of Invasion: An International Multicenter Retrospective Study. *JAMA Otolaryngology-Head & Neck Surgery*. 2014;140(12):1138.
 132. Langius JA, Bakker S, Rietveld DH, Kruizenga HM, Langendijk JA, Weijs PJ, et

- al. Critical weight loss is a major prognostic indicator for disease-specific survival in patients with head and neck cancer receiving radiotherapy. *Br J Cancer*. 2013;109(5):1093-9.
133. Baptistella AR, Hilleshein KD, Beal C, Brambatti JS, Caron R, Baptistella SF, et al. Weight loss as a prognostic factor for recurrence and survival in oropharyngeal squamous cell carcinoma patients. *Mol Clin Oncol*. 2018;9(6):666-72.
 134. Ding W, Liu T, Liang J, Hu T, Cui S, Zou G, et al. Supraglottic squamous cell carcinomas have distinctive clinical features and prognosis based on subregion. *PLoS One*. 2017;12(11):e0188322.
 135. Issa MR, Samuels SE, Bellile E, Shalabi FL, Eisbruch A, Wolf G. Tumor Volumes and Prognosis in Laryngeal Cancer. *Cancers (Basel)*. 2015;7(4):2236-61.
 136. Miyabe J, Hanamoto A, Tatsumi M, Hamasaki T, Takenaka Y, Nakahara S, et al. Metabolic tumor volume of primary tumor predicts survival better than T classification in the larynx preservation approach. *Cancer Sci*. 2017;108(10):2030-8.
 137. Welkoborsky HJ, Hinni M, Dienes HP, Mann WJ. Predicting recurrence and survival in patients with laryngeal cancer by means of DNA cytometry, tumor front grading, and proliferation markers. *Ann Otol Rhinol Laryngol*. 1995;104(7):503-10.
 138. Starska K, Kulig A, Lukomski M. Tumor front grading in prediction of survival and lymph node metastases in patients with laryngeal carcinoma. *Adv Med Sci*. 2006;51:200-4.
 139. Wiernik G, Millard PR, Haybittle JL. The predictive value of histological classification into degrees of differentiation of squamous cell carcinoma of the larynx and hypopharynx compared with the survival of patients. *Histopathology*. 1991;19(5):411-7.
 140. Groome PA, Schulze K, Boysen M, Hall SF, Mackillop WJ. A comparison of published head and neck stage groupings in carcinomas of the oral cavity. *Head & Neck*. 2001;23(8):613-24.
 141. Beesley LJ, Hawkins PG, Amlani LM, Bellile EL, Casper KA, Chinn SB, et al. Individualized survival prediction for patients with oropharyngeal cancer in the human papillomavirus era. *Cancer*. 2019;125(1):68-78.
 142. Ang KK, Harris J, Wheeler R, Weber R, Rosenthal DI, Nguyen-Tan PF, et al. Human papillomavirus and survival of patients with oropharyngeal cancer. *N Engl J Med*. 2010;363(1):24-35.
 143. Matos LL, Miranda GA, Cernea CR. Prevalence of oral and oropharyngeal human papillomavirus infection in Brazilian population studies: a systematic review. *Braz J Otorhinolaryngol*. 2015;81(5):554-67.
 144. Bratman SV, Bruce JP, O'Sullivan B, Pugh TJ, Xu W, Yip KW, et al. Human Papillomavirus Genotype Association With Survival in Head and Neck Squamous Cell Carcinoma. *JAMA Oncol*. 2016;2(6):823-6.
 145. Lajer CB, von Buchwald C. The role of human papillomavirus in head and neck cancer. *Apmis*. 2010;118(6-7):510-9.
 146. Sabin SL, Rosenfeld RM, Sundaram K, Har-el G, Lucente FE. The impact of comorbidity and age on survival with laryngeal cancer. *Ear Nose Throat J*. 1999;78(8):578, 81-4.
 147. Datema FR, Ferrier MB, van der Schroeff MP, Baatenburg de Jong RJ. Impact of comorbidity on short-term mortality and overall survival of head and neck cancer patients. *Head Neck*. 2010;32(6):728-36.
 148. Karadaghy OA, Shew M, New J, Bur AM. Development and Assessment of a Machine Learning Model to Help Predict Survival Among Patients With Oral Squamous Cell Carcinoma. *JAMA Otolaryngol Head Neck Surg*. 2019.

149. Kwok J, Langevin SM, Argiris A, Grandis JR, Gooding WE, Taioli E. The impact of health insurance status on the survival of patients with head and neck cancer. *Cancer*. 2010;116(2):476-85.
150. Lee DH, Kim MJ, Roh JL, Kim SB, Choi SH, Nam SY, et al. Distant metastases and survival prediction in head and neck squamous cell carcinoma. *Otolaryngol Head Neck Surg*. 2012;147(5):870-5.

APÊNDICES

APÊNDICE 1

APROVAÇÃO

O Coordenador da Comissão de Ética para Análise de Projetos de Pesquisa - CAPPesq da Diretoria Clínica do Hospital das Clínicas da Faculdade de Medicina da Universidade de São Paulo, em 03/05/2011, **APROVOU ad-referendum as modificações** abaixo mencionadas referente ao Protocolo de Pesquisa nº **0511/07**, intitulado: "**BUSCA DE MARCADORES DE AGRESSIVIDADE EM TUMORES DE CABEÇA E PESCOÇO**", apresentado pelo Departamento de **CIRURGIA**.

- Sub-projeto intitulado "FATORES AMBIENTAIS, CLÍNICOS, HISTOPATOLÓGICOS E MOLECULARES ASSOCIADOS AO DESENVOLVIMENTO E AO PROGNÓSTICO DE CARCINOMAS EPIDERMÓIDES DE CABEÇA E PESCOÇO"
- Termo de Consentimento Livre e Esclarecido para Controles (Sub-projeto)
- Termo de Consentimento Livre e Esclarecido para Casos (Sub-projeto)
- Questionário inicial sobre hábitos e estilo de vida versão Março/2010
- Formulário de Informações Clínicas versão Março/2010
- Questionário de acompanhamento sobre hábitos e estilo de vida (Somente para os casos) (A ser completado nas visitas anuais após o tratamento) versão Março/2010
- Formulário para seguimento (somente para casos) versão Março/2010

Pesquisador (a) Responsável: **PROF. DR. PEDRO MICHALUART JR.**

CAPPesq, 04 de maio de 2011

PROF. DR. EUCLIDES AYRES DE CASTILHO
Coordenador
Comissão de Ética para Análise de
Projetos de Pesquisa - CAPPesq

Comissão de Ética para Análise de Projetos de Pesquisa do HCFMUSP da Diretoria Clínica do Hospital das Clínicas da Faculdade de Medicina da Universidade de São Paulo Rua Ovídio Pires de Campos, 225, 5º andar - CEP 05403 010 - São Paulo – SP Fone: 011 3069 6442 Fax: 011 3069 6492
e-mail: cappesq@hcnet.usp.br

APÊNDICE 2

Hospital das Clínicas da FMUSP
Comissão de Ética para Análise de Projetos de Pesquisa
CAPPesq

Nº Protocolo: 0511/07

Título: BUSCA DE MARCADORES DE AGRESSIVIDADE EM TUMORES DE CABEÇA E PESCOÇO

Pesquisador Responsável: Raquel Ajub Moyses

Pesquisador Executante: Raquel Ajub Moyses

Disciplina: Cirurgia de Cabeça e Pescoço

Departamento: CIRURGIA

A Comissão de Ética para Análise de Projetos de Pesquisa – CAPPesq da Diretoria Clínica do Hospital das Clínicas da Faculdade de Medicina da Universidade de São Paulo, **APROVOU / TOMOU CIÊNCIA** na sessão datada de 20/01/2016, do(s) documento(s) abaixo mencionado(s) conforme parecer anexo.

Carta datada de 10/12/2015 - Solicitação de inclusão do pesquisador Leonardo Freitas Boaventura Rios no subprojeto intitulado "Modelo de predição de morte em 5 anos para pacientes com CECCP", baseado em dados epidemiológicos, clínicos, histopatológicos e positividade para HPV deste subprojeto.

Em conformidade com a Resolução CNS nº 466/12 – cabe ao pesquisador: **a)** desenvolver o projeto conforme delineado; **b)** elaborar e apresentar relatórios parciais e final; **c)** apresentar dados solicitados pelo CEP, a qualquer momento; **d)** manter em arquivo sob sua guarda, por 5 anos da pesquisa, contendo fichas individuais e todos os demais documentos recomendados pelo CEP; **e)** encaminhar os resultados para publicação, com os devidos créditos aos pesquisadores associados e ao pessoal técnico participante do projeto; **f)** justificar perante ao CEP interrupção do projeto ou a não publicação dos resultados.

CAPPesq, 20 de Janeiro de 2016

PROF. DR. ALFREDO JOSÉ MANSUR

Coordenador

**Comissão de Ética para Análise de
Projetos de Pesquisa - CAPPesq**

APÊNDICE 2 (continuação)

Hospital das Clínicas da FMUSP
Comissão de Ética para Análise de Projetos de Pesquisa
CAPPesq

Ao

Departamento de Cirurgia

A Comissão de Ética para Análise de Projetos de Pesquisa-CAPPesq da Diretoria Clínica do Hospital das Clínicas da Faculdade de Medicina da Universidade de São Paulo, **APROVOU** na sessão de 20.01.16 conforme parecer anexo a execução do subprojeto intitulado **"Modelo de predição de morte para pacientes com carcinoma epidermóide de cabeça e pescoço"**, como tese de doutorado do aluno **Leonardo Freitas Boaventura Rios**.

O referido projeto faz parte do **Protocolo de Pesquisa nº 0511/07**, intitulado **"BUSCA DE MARCADORES DE AGRESSIVIDADE EM TUMORES DE CABEÇA E PESCOÇO"**, aprovado por esta CAPPesq em 01.08.2007.

Pesquisador Responsável: Raquel Ajub Moyses

CAPPesq, 20 de Janeiro de 2016

Prof. Dr. Alfredo José Mansur
Coordenador

Comissão de Ética para Análise de
Projetos de Pesquisa - CAPPesq

APÊNDICE 3

5/22/2017

>>> Clinical Genomics <<<

FORMULÁRIO DE CADASTRO DOS PACIENTES

::Dados de identificação

Q1- Nº de Identificação:

Q2- Nome do Paciente:

Q3- Número do prontuário no hospital:

Q4- Hospital:

Q5- Paciente: ☐ Internado ☐ Ambulatorial

Q6- Diagnóstico principal: ☐ Controle ☐ Câncer

Q7- CID-10(código): Q8- CID-10(Descrição da doença):

Q9 - Entrevistador:

::Informações Gerais

Q10- Data da Entrevista: dd/mm/aaaa

Q11- Início da entrevista: hh:mm

Q12- Coleta de Sangue: ☐ Sim ☐ Não Q13- Data: dd/mm/aaaa

Q14 - Sexo: ☐ Feminino ☐ Masculino

Meu nome é De início eu quero agradecer o(a) senhor(a) por participar nesse estudo. Nós estamos conduzindo um estudo com a finalidade de esclarecer se determinadas características e hábitos de homens e mulheres podem ter relação com algumas doenças. Eu farei várias perguntas cujas respostas serão registradas neste caderno. Devo dizer que tudo que o(a) senhor(a) responder na entrevista será estritamente confidencial e as informações colhidas das inúmeras pessoas que irão participar do estudo serão usadas apenas em relatos científicos, sem nenhuma identificação pessoal. Os possíveis benefícios desse estudo dependem de que as respostas sejam as mais reais (verdadeiras, sinceras) possíveis. Por favor pergunte se não entender o significado de alguma questão. A qualquer momento o(a) senhor(a) pode se recusar a continuar ou a responder perguntas específicas. Além do questionário o estudo inclui a coleta de uma amostra de sangue.

Se houver necessidade de entrar em contato com o(a) senhor(a), poderia fornecer seu endereço e telefone?

Q15 - Endereço:

Q16 - Bairro:

Q17 - Cidade: Q18 - UF da cidade: Q19 - descrição da cidade:

Q20 - Cep: 00000-000

Q21 - Telefone: (DDD)- 0000-0000

Podemos começar? O(A) Sr(a) poderia assinar essa folha de consentimento?

Q22 - Data de Nascimento: dd/mm/aaaa

Q23 - Qual é a sua idade?

Qual a sua profissão? (aquela que o(a) sr(a) exerceu por mais tempo): Q24- Q25- Descrição: (CBO)

Q26 - Por quanto tempo o(a) sr(a) exerceu esta profissão? anos

Pergunte ao entrevistado como ele/ela se classifica racialmente. Pergunte sobre a raça/origem dos avós. O entrevistador decide sobre a classificação racial.

Q27- Raça: ☐ Branco ☐ Mulato ☐ Negro ☐ Oriental ☐ Índio Q28-(outraraça) ☐ Outra:

Q29 - O(A) Sr.(a) já fez algum tratamento para a doença atual? (A doença que o(a) trouxe ao hospital agora?) ☐ Sim ☐ Não

Q30 - Se "Sim", então qual tratamento?

Q31 - Qual a cidade onde o(a) Sr.(a) mora? Q32 - UF da cidade: Q33 - descrição da cidade:

Q34- Há quanto tempo o(a) Sr.(a) mora nessa cidade?(anos) Se menos de 1 ano codifique como 00

Se o(a) Sr.(a) esta morando há menos de 1 ano nessa cidade, onde o(a) Sr.(a) morava antes?

5/22/2017

>>> Clinical Genomics <<<

Q35 - Q36 - UF da cidade: Q37 - descrição da cidade:

Em que cidade/estado/país o(a) sr.(a) nasceu?

Q38 (nacionalidade) - ☐ Brasileiro ☐ Estrangeiro Q39(onde nasceu) - Q40 - UF da cidade: Q41 - descrição da cidade: Q42(onde nasceu) - País: Q43 - O(A) sr.(a) estudou em escola? ☐ Sim ☐ NãoQ44 - Selecione a opção mais elevada quanto ao grau de instrução: ☐ Alfabetizado ☐ Primário Completo☐ Primário Incompleto ☐ Secundário Completo ☐ Secundário Incompleto ☐ Técnico ☐ Universitário**::História de Câncer no Entrevistado e na Família**Quanto irmãos o(a) Sr(a) teve? : Q45 - Quanto irmãos o(a) Sr(a) teve? : Q46 - Quanto filhos o(a) Sr(a) teve? : Q47 - Quanto filhas o(a) Sr(a) teve? : Q48 - **OBS:** Segue as opções de preenchimento : Coloque os números e letras em vermelho correspondentes aos campos nos quadros em branco da tabela anexada abaixo :

Tipo de familiar : Próprio(a) Paciente : 01 Mãe: 02 Pai: 03 Irmão: 04 Irmã : 05 Filho: 06 Filha : 07

Vivo/Morto : Vivo: V Morto: M

Tumor : Sim: S Não: N IGN: I

Tabagismo : Sim: S Não: N IGN: I

	C1	C2	C3	C4	C5	C6	C7
T1	Tipo de familiar	Vivo/Morto	Idade atual ou da morte	Tumor	Tipo de tumor (CID-10)	Idade no diagnóstico	Tabagista
L1	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>
L2	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>
L3	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>
L4	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>
L5	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>
L6	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>
L7	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>
L8	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>
L9	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>
L10	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>
L11	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>
L12	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>
L13	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>
L14	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>
L15	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>
L16	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>
L17	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>
L18	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>
L19	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>
L20	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>
Ln	T1_LnC1	T1_LnC2	T1_LnC3	T1_LnC4	T1_LnC5	T1_LnC6	T1_LnC7

OBS: No banco de dados esta tabela acima está codificada da seguinte forma: Tx_LyCz, onde "x", "y" e "z" representam, respectivamente, "x"

o número da tabela, "y" o número das novas informações incluídas e "z" o número do campo de informações (tipo de familiar:vivo/morto,idade atual ou da morte,tipo do tumor,idade no diagnóstico ou tabagista). Desta forma, o Ly é a localização de uma nova linha e o Cz é a localização da coluna na tabela.

Ex: T1_L3C2 - Tabela1 apresentando informações da linha 3 (y=3) da tabela de história familiar, e da 2ª coluna (C2), que representa o campo com informações de Vivo/Morto.

:.História de Tabagismo

Q49 - O(A) Sr.(a) fuma ou já fumou em média 1 cigarro, charuto ou cachimbo, diariamente, pelo menos por 1 ano?

☐ Sim, ainda fuma ☐ Nunca fumou ☐ Somente no passado

(se parou de fumar há menos de 12 meses assinalar(sim, ainda fuma))

Gostaríamos que o(a) sr.(a) descrevesse os períodos de sua vida em que fumou cigarro, charuto ou cachimbo, as quantidades que fumou e outros detalhes. Por favor, tente lembrar as mudanças mais importantes quanto à quantidade e tipo de cigarro. Ignore as mudanças que ocorreram por períodos curtos(menos de 1 ano).

Nota para o entrevistador: Evite a superposição de anos para consumo do mesmo tipo de cigarro, por exemplo: utilize 30-40, 41-45 ao invés de 30-40, 40-45.

OBS: Segue as opções de preenchimento : Coloque os números em vermelho correspondentes aos campos nos quadros em branco da tabela anexada abaixo : **Tipo de Tabaco :** Cigarro : **01** Charuto : **02** Cachimbo : **03** Cigarro de palha: **04**

Que tipo de tabaco o(a) sr.(a) fuma ou já fumou?

	C1	C2	C3	C4
T2	Tipo de Tabaco	Idade de Início	Idade que Parou	Nº por Dia
L1	☐	☐	☐	☐
L2	☐	☐	☐	☐
L3	☐	☐	☐	☐
L4	☐	☐	☐	☐
L5	☐	☐	☐	☐
Ln	T2_LnC1	T2_LnC2	T2_LnC3	T2_LnC4

OBS: No banco de dados esta tabela acima está codificada da seguinte forma: Tx_LyCz, onde "x", "y" e "z" representam, respectivamente, "x" o número da tabela, "y" o número das novas informações incluídas e "z" o número do campo de informações (tipo de tabaco,idade de início, idade que parou, nº por dia). Desta forma, o Ly é a localização de uma nova linha e o Cz é a localização da coluna na tabela. Ex: T2_L3C2 -

Tabela2 apresentando informações da linha 3 (y=3) da tabela de tabaco, e da 2ª coluna (C2), que representa o campo com informações de idade de início.

:.Hábitos de Bebida

Q50 - O(A) Sr(a) já tomou (ingeriu) bebidas alcoólicas pelo menos 1 vez por mês?

☐ Sim, ainda bebe ☐ Nunca ☐ Só no passado

(se parou nos últimos 12 meses, a resposta é (sim ainda bebe)

Q51 - Quando o(a) Sr(a) costuma beber(ou bebia)?

☐ Nas refeições ☐ Entre as refeições ☐ Aos finais de semana ☐ Ambos

Descreva os períodos de sua vida durante os quais o(a) tomou bebidas alcoólicas. Por favor, tente resumir as mudanças mais importantes em sua vida em relação a quantidade e tipo de bebida. Ignore quaisquer mudanças ocorridas durante curtos períodos de tempo(menos de 1 ano), ou bebidas consumidas ocasionalmente.

NOTA PARA O ENTREVISTADOR: Evite a superposição de anos de consumo de uma mesma bebida, por exemplo: utilize 30-40,41-45ao invés de 30-40,40-45, perguntar separadamente para cada bebida.

OBS. Destilados: uísque, conhaque, cachaça e vodca.

OBS: Segue as opções de preenchimento : Coloque os números em vermelho correspondentes aos campos nos quadros em branco da tabela anexada abaixo :

Bebida : Vinho : **01** Cerveja : **02** Destilados : **03**

Unidade: Copo pequeno -50ml: **01** Copo médio -100ml: **02** Copo grande -250ml: **03** 1/2 garrafa pequena ou lata - 330ml: **04**

Garrafa 700-750ml : 05 Garrafa - 1L : 06

Por : Dia : **01** Semana : **02** Mês : **03**

	C1	C2	C3	C4	C5	C6
T3	Bebida	Idade de Início	Idade que Parou	Unidade	Unidade Consumidas	Por
L1	☐	☐	☐	☐	☐	☐
L2	☐	☐	☐	☐	☐	☐
L3	☐	☐	☐	☐	☐	☐
L4	☐	☐	☐	☐	☐	☐
L5	☐	☐	☐	☐	☐	☐
Ln	T3_LnC1	T3_LnC2	T3_LnC3	T3_LnC4	T3_LnC5	T3_LnC6

5/22/2017

>>> Clinical Genomics <<<

OBS: No banco de dados esta tabela acima está codificada da seguinte forma: Tx_LyCz, onde "x", "y" e "z" representam, respectivamente, "x" o número da tabela, "y" o número das novas informações incluídas e "z" o número do campo de informações (tipo de bebida, idade de início, idade que parou, unidade, unidade consumida e por). Desta forma, o Ly é a localização de uma nova linha e o Cz é a localização da coluna na tabela. Ex: T3_L3C4
- Tabela3 apresentando informações da linha 3 (y=3) da tabela de bebida, e da 4ª coluna (C4) que representa o campo com informações de unidade.

APÊNDICE 4

5/22/2017

>>> Clinical Genomics <<<

FORMULÁRIO DE DADOS CLÍNICOS E LABORATORIAIS: CABEÇA E PESCOÇO

Q0 - N° de Identificação: Nome do Paciente:

::Dados Clínicos e Laboratoriais ao Diagnóstico de Cabeça e Pescoço

PRIMEIRA CONSULTA

QUEIXA E DURAÇÃO

Q201- Nome da Mãe:

Q1 - Peso habitual antes da doença: (KG)

Q2 - Altura: (cm)

Q3 - Perda de Peso: ☐ Sim ☐ Não Q4 - (KG) Q5 - (%) Q6 - (meses)

Q7 - IMC:

Q8 - Alterações Visuais: ☐ Sim ☐ Não Q9 - (dias)

Q10 - Dificuldade Respiratória/Dispneia: ☐ Sim ☐ Não Q11 - (dias)

Q12 - Traqueostomia: ☐ Sim ☐ Não - Q13 - Data Realização: dd/mm/aaaa

Q14 - Disfagia: ☐ Sim ☐ Não Q15 - (dias)

Q16 - Disfonia: ☐ Sim ☐ Não Q17 - (dias)

Q18 - Dor: ☐ Sim ☐ Não Q19 - (dias)

Q20 - Feridas: ☐ Sim ☐ Não Q21 - (dias) Q22 - Localização:

Q23 - Lesões externas visíveis: ☐ Sim ☐ Não Q24 - (dias) Q25 - Localização:

Q26 - Nódulos ou massas: ☐ Sim ☐ Não Q27 - (dias) Q28 - Localização:

Q29 - Obstrução Nasal: ☐ Sim ☐ Não Q30 - (dias) Q31 - Lado: ☐ Esquerda ☐ Direita

Q32 - Odinofagia: ☐ Sim ☐ Não Q33 - (dias)

Q34 - Otalgia: ☐ Sim ☐ Não Q35 - (dias)

Q36 - Rinorréia: ☐ Sim ☐ Não Q37 - (dias)

Q38 - Sangramentos: ☐ Sim ☐ Não Q39 - (dias)

Q40 - Sensação de corpo estranho: ☐ Sim ☐ Não Q41 - (dias) Q42 - Localização:

Q43 - Tosse: ☐ Sim ☐ Não Q44 - (dias) Q45 - Outras Queixas:

LOCALIZAÇÃO DA LESÃO

CID-10

- Q46 - ☐ C00 - Neoplasia maligna do lábio
- Q47 - ☐ C00.0 - Lábio superior externo
- Q48 - ☐ C00.1 - Lábio inferior externo
- Q49 - ☐ C00.2 - Lábio externo, não especificado
- Q50 - ☐ C00.3 - Lábio superior, face interna
- Q51 - ☐ C00.4 - Lábio inferior, face interna
- Q52 - ☐ C00.5 - Lábio, sem especificação, face interna
- Q53 - ☐ C00.6 - Comissura labial
- Q54 - ☐ C00.8 - Lesão invasiva do lábio
- Q55 - ☐ C00.9 - Lábio, não especificado

- Q56 - ☐ C01 - Neoplasia maligna de base de língua
- Q57 - ☐ C02 - Neoplasia maligna de outras partes e partes não especificadas de língua
- Q58 - ☐ C02.0 - Face dorsal da língua
- Q59 - ☐ C02.1 - Borda da língua
- Q60 - ☐ C02.2 - Face ventral da língua

- ☐ Q61 - ☐ C02.3 - Dois terços anteriores da língua, parte não especificada
- ☐ Q62 - ☐ C02.4 - Amígdala lingual
- ☐ Q63 - ☐ C02.8 - Lesão invasiva da língua
- ☐ Q64 - ☐ C02.9 - Língua, não especificada
- ☐ Q65 - ☐ C03 - Neoplasia maligna de gengiva
- ☐ Q66 - ☐ C03.0 - Gengiva superior
- ☐ Q67 - ☐ C03.1 - Gengiva inferior
- ☐ Q68 - ☐ C03.9 - Gengiva, não especificada
- ☐ Q69 - ☐ C04 - Neoplasia maligna de assoalho da boca
- ☐ Q70 - ☐ C04.0 - Assoalho anterior da boca
- ☐ Q71 - ☐ C04.1 - Assoalho lateral da boca
- ☐ Q72 - ☐ C04.8 - Lesão invasiva do assoalho da boca
- ☐ Q73 - ☐ C04.9 - Assoalho da boca, não especificado
- ☐ Q74 - ☐ C05 - Neoplasia maligna de palato
- ☐ Q75 - ☐ C05.0 - Palato duro
- ☐ Q76 - ☐ C05.1 - Palato mole
- ☐ Q77 - ☐ C05.2 - Úvula
- ☐ Q78 - ☐ C05.8 - Lesão invasiva do palato
- ☐ Q79 - ☐ C05.9 - Palato não especificado
- ☐ Q80 - ☐ C06 - Neoplasia maligna de outras partes e de partes não especificadas da boca
- ☐ Q81 - ☐ C06.0 - Mucosa oral
- ☐ Q82 - ☐ C06.1 - Vestíbulo da boca
- ☐ Q83 - ☐ C06.2 - Área retromolar
- ☐ Q84 - ☐ C06.8 - Lesão invasiva de outras partes e de partes não especificadas da boca
- ☐ Q85 - ☐ C06.9 - Boca, não especificada
- ☐ Q86 - ☐ C09 - Neoplasia maligna da amígdala
- ☐ Q87 - ☐ C09.0 - Fossa amigdalana
- ☐ Q88 - ☐ C09.1 - Pilar amigdalano(anterior)(posterior)
- ☐ Q89 - ☐ C09.8 - Lesão invasiva da amígdala
- ☐ Q90 - ☐ C09.9 - Amígdala, não especificada
- ☐ Q91 - ☐ C10 - Neoplasia maligna da orofaringe
- ☐ Q92 - ☐ C10.0 - Valécula
- ☐ Q93 - ☐ C10.1 - Face anterior da epiglote
- ☐ Q94 - ☐ C10.2 - Parede lateral da orofaringe
- ☐ Q95 - ☐ C10.3 - Parede posterior da orofaringe
- ☐ Q96 - ☐ C10.4 - Fenda branquial
- ☐ Q97 - ☐ C10.8 - Lesão invasiva da orofaringe
- ☐ Q98 - ☐ C10.9 - Orofaringe, não especificada
- ☐ Q99 - ☐ C11 - Neoplasia maligna da nasofaringe
- ☐ Q100 - ☐ C11.0 - Parede superior da nasofaringe
- ☐ Q101 - ☐ C11.1 - Parede posterior da nasofaringe
- ☐ Q102 - ☐ C11.2 - Parede lateral da nasofaringe
- ☐ Q103 - ☐ C11.3 - Parede anterior da nasofaringe
- ☐ Q104 - ☐ C11.8 - Lesão invasiva da nasofaringe
- ☐ Q105 - ☐ C11.9 - Nasofaringe, não especificada

- Q106 - ☐ C12 - Neoplasia maligna do seio piriforme
- Q107 - ☐ C13 - Neoplasia maligna da hipofaringe
- Q108 - ☐ C13.0 - Região pós-cricóide
- Q109 - ☐ C13.1 - Prega ariepiglótica, face hipofaríngea
- Q110 - ☐ C13.2 - Parede posterior da hipofaringe
- Q111 - ☐ C13.8 - Lesão invasiva da orofaringe
- Q112 - ☐ C13.9 - Hipofaringe, não especificada
- Q113 - ☐ C14 - Neoplasia maligna de outras localidades e de localidades mal definidas do lábio cavidade oral e faringe
- Q114 - ☐ C14.0 - Faringe, não especificada
- Q115 - ☐ C14.1 - Laringofaringe, descrita como tal
- Q116 - ☐ C14.2 - Anel de Waldeyer
- Q117 - ☐ C14.8 - Lesão invasiva do lábio, cavidade oral e faringe
- Q118 - ☐ C30 - Neoplasia maligna da cavidade nasal e do ouvido médio
- Q119 - ☐ C30.0 - Cavidade nasal
- Q120 - ☐ C30.1 - Ouvido médio
- Q121 - ☐ C31 - Neoplasia maligna dos seios da face
- Q122 - ☐ C31.0 - Seio maxilar
- Q123 - ☐ C31.1 - Seio etmoidal
- Q124 - ☐ C31.2 - Seio frontal
- Q125 - ☐ C31.3 - Seio esfenoidal
- Q126 - ☐ C31.8 - Lesão invasiva dos seios da face
- Q127 - ☐ C31.9 - Seio da face, não especificado
- Q128 - ☐ C32 - Neoplasia maligna da laringe
- Q129 - ☐ C32.0 - Glote
- Q130 - ☐ C32.1 - Região supraglótica
- Q131 - ☐ C32.2 - Região subglótica
- Q132 - ☐ C32.3 - Cartilagens da laringe
- Q133 - ☐ C32.8 - Lesão invasiva da laringe
- Q134 - ☐ C32.9 - Laringe não especificada
- Q135 - ☐ C77.0 - Neoplasia maligna secundária e não especificada dos gânglios linfáticos
- Q136 - ☐ C44.9 - Neoplasia maligna da pele

TAMANHO DO TUMOR

Tamanho: Q137 - cm X Q138 - cm X Q139 - cm

LINFONOS COMPROMETIDOS CLINICAMENTE

Q140 - E1 NIVEL I D: ☐ Sim ☐ Não Q141 - Nº de Linfonodos:

Q143 - E2 NIVEL II D: ☐ Sim ☐ Não Q144 - Nº de Linfonodos:

Q146 - E3 NIVEL III D: ☐ Sim ☐ Não Q147 - Nº de Linfonodos:

Q149 - E4 NIVEL IV D: ☐ Sim ☐ Não Q150 - Nº de Linfonodos:

Q142 - Maior Linfonodo: cm

Q145 - Maior Linfonodo: cm

Q148 - Maior Linfonodo: cm

Q151 - Maior Linfonodo: cm

5/22/2017

>>> Clinical Genomics <<<

☐ Sim ☐ Não	☐ Sim ☐ Não	☐ Sim ☐ Não	☐ Sim ☐ Não
Q152 -E5 NIVEL V D:	☐ Sim ☐ Não	Q153 -Nº de Linfonodos:	☐ Sim ☐ Não
Q155 -E6 NIVEL I E:	☐ Sim ☐ Não	Q156 -Nº de Linfonodos:	☐ Sim ☐ Não
Q158 -E7 NIVEL II E:	☐ Sim ☐ Não	Q159 -Nº de Linfonodos:	☐ Sim ☐ Não
Q161 -E8 NIVEL III E:	☐ Sim ☐ Não	Q162 -Nº de Linfonodos:	☐ Sim ☐ Não
Q164 -E9 NIVEL IV E:	☐ Sim ☐ Não	Q165 -Nº de Linfonodos:	☐ Sim ☐ Não
Q167 -E10 NIVEL V E:	☐ Sim ☐ Não	Q168 -Nº de Linfonodos:	☐ Sim ☐ Não
Q170 -E11 NIVEL VI:	☐ Sim ☐ Não	Q171 -Nº de Linfonodos:	☐ Sim ☐ Não
Q173 -E12 NIVEL VII:	☐ Sim ☐ Não	Q174 -Nº de Linfonodos:	☐ Sim ☐ Não

METÁSTASES À DISTÂNCIA

Q176 -SNC:	☐ Sim ☐ Não	Q177 -Data:	<input type="text"/> dd/mm/aaaa	Q154 -Maior Linfonodo:	<input type="text"/> cm
Q178 -Pulmão:	☐ Sim ☐ Não	Q179 -Data:	<input type="text"/> dd/mm/aaaa	Q157 -Maior Linfonodo:	<input type="text"/> cm
Q190 -Fígado:	☐ Sim ☐ Não	Q181 -Data:	<input type="text"/> dd/mm/aaaa	Q160 -Maior Linfonodo:	<input type="text"/> cm
Q182 -Ossos:	☐ Sim ☐ Não	Q183 -Data:	<input type="text"/> dd/mm/aaaa	Q163 -Maior Linfonodo:	<input type="text"/> cm
Q195 -Outros órgãos:	☐ Sim ☐ Não	Q186 -Data:	<input type="text"/> dd/mm/aaaa	Q166 -Maior Linfonodo:	<input type="text"/> cm
				Q169 -Maior Linfonodo:	<input type="text"/> cm
				Q172 -Maior Linfonodo:	<input type="text"/> cm
				Q175 -Maior Linfonodo:	<input type="text"/> cm

TMN Clínico: Q188 -T ☐ Q189 -N ☐ Q190 -M ☐

INFORMAÇÕES TERAPÊUTICAS

Q191 -Cirurgia realizada:	☐ Sim ☐ Não	Q192 -Data:	<input type="text"/> dd/mm/aaaa	Q184 -Qual ?:	<input type="text"/>
Q193 - Notas:					
<input type="text"/>					
Q194 -Radioterapia:	☐ Sim ☐ Não	Q195 -Data início:	<input type="text"/> dd/mm/aaaa	Q196 -Data término:	<input type="text"/>
Q197 -Tipo de Radioterapia:	<input type="text"/>				
Q198 -Dose:	<input type="text"/> cGy				
Q199 -Quimioterapia:	☐ Sim ☐ Não				

(Preencha a tabela abaixo somente se a sua resposta em Quimioterapia for Sim)

	C1	C2	C3	C4	C5	C6
T1	Drogas	Dose	Nº Dias	Ciclos	Data Int.	Data Fim
L1	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>
L2	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>
L3	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>
L4	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>
L5	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>
Ln	T1_LnC1	T1_LnC2	T1_LnC3	T1_LnC4	T1_LnC5	T1_LnC6

OBS: No banco de dados esta tabela acima está codificada da seguinte forma: Tx_LyCz, onde "x", "y" e "z" representam, respectivamente, "x" o número da tabela, "y" o número das novas informações incluídas e "z" o número do campo de informações (drogas,dose,nº dias,ciclos, data início ou data fim). Desta forma, o Ly é a localização de uma nova linha e o Cz é a localização da coluna na tabela. Ex: T1_L3C2 - Tabela1 apresentando informações da linha 3 (y=3) da tabela de quimioterapia, e da 2ª coluna (C2), que representa o campo com informações de dose.

Q200 -Tipo de Quimioterapia (especificar):

APÊNDICE 5

5/22/2017

>>> Clinical Genomics <<<

FORMULÁRIO ANATOMO PATOLÓGICO: CABEÇA E PESCOÇO

OBS: Ao utilizar o banco de dados relativo a este formulário, deve-se verificar que há um código formado pela letra "A" seguido de um número, antes do código de cada questão. Como para um mesmo paciente pode haver mais de um anátomo patológico, este código inicial com a letra "A" indica o número do formulário de anátomo patológico inserido no sistema. Suponha o exemplo A2_Q3: "A2" indica o o segundo formulário de anátomo patológico e "Q3" a terceira questão deste formulário.

Q0 - Nº de Identificação Nome do Paciente:

1. Protocolo Anátomo Patológico Cabeça e Pescoço

1) TIPO DE MATERIAL

Nota: Preencher o campo biópsia e as informações subsequentes somente para os casos em que o paciente não foi submetido a cirurgia.

Q1 - ☐ Biópsia ☐ Peça Cirúrgica - Q2 - data do exame: dd/mm/aaaa

2) EXAME MACROSCÓPICO

Tamanho do Tumor - dimensão máxima em centímetros (maiores diâmetros e espessura tomada a partir da superfície mucosa normal)

Q3 cm X Q4 cm X Q5 cm Q6 - ☐ Não-avaliável

Q7 - Crescimento:

☐ Exofítico ☐ Endofítico ☐ Não-avaliável

Q8 - Área de comprometimento do tumor

3) EXAME MICROSCÓPICO

Tipo Histológico - Carcinoma Epidermóide

Q9 - Carcinoma espinho celular (CEC): ☐ Sim ☐ Não ☐ Não Informado

Q9 - Variante Morfológica:

☐ Verrucoso ☐ Adenoescamoso
☐ Linfopitelial ☐ Escamoso Basalóide ☐ Escamoso Adenóide ☐ Fusocelular (Sarcomatóide)
☐ Não-avaliável ☐ Outro

Q10

Q11 - Grau de diferenciação:

☐ Bem diferenciado ☐ Moderadamente ☐ Pouco Diferenciado ☐ Não-avaliável

Para microinvasivo (o componente in situ predomina, com invasão focal da lâmina própria)

☐ Microinvasivo ☐ Lâmina Própria ☐ Plano Muscular ☐ Cartilagem ☐ Outras Estruturas

Q12 - Profundidade de invasão do tumor:

☐ Não-avaliável

Infiltração Vascular

Q13 - Sanguínea:

☐ Ausente ☐ Presente ☐ Não-avaliável

Q14 - Linfática:

☐ Ausente ☐ Presente ☐ Não-avaliável

Q15 - Invasão Peri-neural:

☐ Ausente ☐ Presente ☐ Não-avaliável

Q16 - Infiltrado inflamatório peri-tumoral:

☐ Ausente ☐ Escasso ☐ Moderado ☐ Intenso ☐ Não-avaliável

Q17 - Margens Cirúrgicas:

☐ Com displasia de alto grau (especificar o local):
☐ Envolvida por tumor invasivo (especificar o local):
☐ Com carcinoma in situ (especificar o local):
☐ Não avaliável
☐ Livres de tumor

Q18
Q18
Q18
Q18

Q19 - Lesões em órgãos adjacentes

Linfonodos

NÍVEL I D

Q20 - Nº dissecado

Q21 - nº comprometido

Q22 - invasão extracapsular

NÍVEL II D

Q23 - Nº dissecado

Q24 - nº comprometido

Q25 - invasão extracapsular

NÍVEL III D

Q26 - Nº dissecado

Q27 - nº comprometido

Q28 - invasão extracapsular

NÍVEL IV D

Q29 - Nº dissecado

Q30 - nº comprometido

Q31 - invasão extracapsular

NÍVEL V D

Q32 - Nº dissecado

Q33 - nº comprometido

Q34 - invasão extracapsular

NÍVEL I E

Q35 - Nº dissecado

Q36 - nº comprometido

Q37 - invasão extracapsular

NÍVEL II E

Q38 - Nº dissecado

Q39 - nº comprometido

Q40 - invasão extracapsular

NÍVEL III E

Q41 - Nº dissecado

Q42 - nº comprometido

Q43 - invasão extracapsular

NÍVEL IV E

Q44 - Nº dissecado

Q45 - nº comprometido

Q46 - invasão extracapsular

NÍVEL V E

Q47 - Nº dissecado

Q48 - nº comprometido

Q49 - invasão extracapsular

NÍVEL VI

5/22/2017

>>> Clinical Genomics <<<

Q50-Nº dissecado

Q51-nº comprometido

Q52-Invasão extracapsular

NÍVEL VII

Q53-Nº dissecado

Q54-nº comprometido

Q55-Invasão extracapsular

Estadiamento Patológico:

Q56-T cm X Q57-N cm X Q58-M cm Q59: ☐ Não-avaliável

APÊNDICE 6

5/22/2017

>>> Clinical Genomics <<<

FORMULÁRIO DE ACOMPANHAMENTO: CABEÇA E PESCOÇO

OBS: Ao utilizar o banco de dados relativo a este formulário, deve-se verificar que há um código formado pela letra "A" seguido de um número, antes do código de cada questão. Como para um mesmo paciente pode haver mais de um acompanhamento, este código inicial com a letra "A" indica o número do formulário de acompanhamento inserido no sistema. Suponha o exemplo A2_Q3: "A2" indica o o segundo formulário de acompanhamento e "Q3" a terceira questão deste formulário.

Q0 - Nº de Identificação: Nome do Paciente:

Obs: Para cada retorno do paciente ao acompanhamento, deverá ser aberto um novo questionário de acompanhamento.

::Acompanhamento de Cabeça e Pescoço

Q1- ☐ Sem evidências de doenças - Q2- Data: dd/mm/aaaa

Q3- ☐ Recidiva loco-regional ou residual

Q4- ☐ Metástase à distância

Q5- ☐ 2º Tumor primário

Q6- ☐ Óbito

Q7- ☐ Paciente perdido do acompanhamento - Q8- Data: dd/mm/aaaa

Q9- ☐ Outras situações Q10- Data: dd/mm/aaaa

Se Sim, qual? Q11- ☐ Acompanhamento em outro hospital ☐ Mudou de cidade ☐ Outra - Qual? Q12-

Comentários referente aos campos mencionados acima:

Sem evidências de doenças - Ao clicar neste tópico será aberto somente o campo de inclusão de data de retorno dos paciente, ocultando os demais itens do questionário;

Recidiva loco-regional ou residual - Ao clicar neste tópico será aberto somente os itens A e B deste questionário, demais serão ocultados;

Metástase à distância - Ao clicar neste tópico será aberto somente os itens C e D deste questionário, demais serão ocultados;

2º Tumor primário - Ao clicar neste tópico será aberto somente os itens E e F deste questionário, demais serão ocultados;

Óbito - Ao clicar neste tópico será aberto somente o item G deste questionário, demais serão ocultados.

Paciente perdido do acompanhamento - Ao clicar neste tópico será aberto somente o campo de inclusão de data. Incluir tela de alerta: "Esta opção somente deverá ser escolhida na condição da busca de informações sobre o paciente terem sido esgotadas".

OBS: Se mais de uma opção for escolhida, serão abertos na tela os tópicos relacionados ao itens da opção: Ex: Paciente retomou para acompanhamento, mas teve uma recidiva e veio a óbito, os tópicos a serem marcados serão: ☒ Recidiva loco-regional ou residual ☒ Óbito, esta propriedade não se aplica para os casos da opção escolhida ser: "Paciente Perdido".

A) RECIDIVAS - Recidiva loco-regional ou residual

Q13- Tumor Residual (tumor primário): ☐ Sim ☐ Não

Q14- Data: dd/mm/aaaa

Q15- Recidiva Local: ☐ Sim ☐ Não

Q16- Data: dd/mm/aaaa

Q17- Tumor Residual (doença regional): ☐ Sim ☐ Não

Q18- Data: dd/mm/aaaa

Q19- Recidiva Regional - Cervical: ☐ Sim ☐ Não

Q20- Data: dd/mm/aaaa

Q21- Nível: Q22- ☐ Direito ☐ Esquerdo

Q23- Nível: Q24- ☐ Direito ☐ Esquerdo

Q25- Nível: Q26- ☐ Direito ☐ Esquerdo

B) TRATAMENTO REALIZADO PARA A RECIDIVA - Recidiva loco-regional ou residual

Q27- Notas:

Q28- Cirurgia: ☐ Sim ☐ Não

Q29- Se sim, Data: dd/mm/aaaa

5/22/2017

>>> Clinical Genomics <<<

Q30-Notas:

Q31-Radioterapia: ☐ Sim ☐ NãoQ32- Se sim, Data Início: dd/mm/aaaaQ33- Data Fim:

Q34- Se sim, tipo de Radioterapia:

Q35- Se sim, Dose:

Q36-Quimioterapia: ☐ Sim ☐ Não

T1	C1 Drogas	C2 Dose	C3 Ciclos	C4 Data Início	C5 Data Término
L1	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>
L2	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>
L3	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>
L4	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>
L5	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>
L6	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>
L7	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>
L8	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>
Ln	T1_LnC1	T1_LnC2	T1_LnC3	T1_LnC4	T1_LnC5

OBS: No banco de dados esta tabela acima está codificada da seguinte forma: Tx_LyCz, onde "x", "y" e "z" representam, respectivamente, "x" o número da tabela, "y" o número das novas informações incluídas e "z" o número do campo de informações (drogas,dose,nº dias,ciclos, data início ou data fim). Desta forma, o Ly é a localização de uma nova linha e o Cz é a localização da coluna na tabela. Ex: T1_L3C2 - Tabela1 apresentando informações da linha 3 (y=3) da tabela de quimioterapia da recidiva, e da 2ª coluna (C2), que representa o campo com informações de dose.

Q37 -Tipo de Quimioterapia (especificar):

C) METÁSTASES À DISTÂNCIA

Q38 -SNC: ☐ Sim ☐ NãoQ39-Data: dd/mm/aaaaQ40-Pulmão: ☐ Sim ☐ NãoQ41-Data: dd/mm/aaaaQ42-Fígado: ☐ Sim ☐ NãoQ43-Data: dd/mm/aaaaQ44-Ossos: ☐ Sim ☐ NãoQ45-Data: dd/mm/aaaaQ46- Qual ? : Q47-Outros órgãos: ☐ Sim ☐ NãoQ48-Data: dd/mm/aaaaQ49- Qual ? :

D) TRATAMENTO REALIZADO PARA A METÁSTASES À DISTÂNCIA

Q50-Notas:

Q51-Cirurgia: ☐ Sim ☐ NãoQ52- Se sim, Data: dd/mm/aaaa

Q53-Notas:

Q54-Radioterapia: ☐ Sim ☐ NãoQ55- Se sim, Data Início: dd/mm/aaaaQ56- Data Fim:

Q57- Se sim, tipo de Radioterapia:

Q58- Se sim, Dose:

Q59-Quimioterapia: ☐ Sim ☐ Não

C1 C2 C3 C4 C5

5/22/2017

>>> Clinical Genomics <<<

T2	Drogas	Dose	Ciclos	Data Início	Data Término
L1					
L2					
L3					
L4					
L5					
L6					
L7					
L8					
Ln	T2_LnC1	T2_LnC2	T2_LnC3	T2_LnC4	T2_LnC5

OBS: No banco de dados esta tabela acima está codificada da seguinte forma: Tx_LyCz, onde "x", "y" e "z" representam, respectivamente, "x" o número da tabela, "y" o número das novas informações incluídas e "z" o número do campo de informações (drogas,dose,nº dias,ciclos, data início ou data fim). Desta forma, o Ly é a localização de uma nova linha e o Cz é a localização da coluna na tabela. Ex: T2_L3C2 - Tabela2 apresentando informações da linha 3 (y=3) da tabela de quimioterapia de metástase, e da 2ª coluna (C2), que representa o campo com informações de dose.

Q60-Tipo de Quimioterapia (especificar):

E) SEGUNDA LESÃO - 2º Tumor primário

Q61-Cavidade oral: ☐ Sim ☐ NãoQ62-Data: dd/mm/aaaaQ63-Orofaringe: ☐ Sim ☐ NãoQ64-Data: dd/mm/aaaaQ65-Hipofaringe: ☐ Sim ☐ NãoQ66-Data: dd/mm/aaaaQ67-Laringe: ☐ Sim ☐ NãoQ68-Data: dd/mm/aaaaQ69-Esôfago: ☐ Sim ☐ NãoQ70-Data: dd/mm/aaaaQ71-Pulmão: ☐ Sim ☐ NãoQ72-Data: dd/mm/aaaaQ73-Outros: ☐ Sim ☐ NãoQ74-Data: dd/mm/aaaaQ75-Qual?:

F) TRATAMENTO DA SEGUNDA LESÃO - 2º Tumor primário

Q76-Notas:

Q77-Cirurgia: ☐ Sim ☐ NãoQ78-Se sim, Data: dd/mm/aaaa

Q79-Notas:

Q80-Radioterapia: ☐ Sim ☐ NãoQ81-Se sim, Data Início: dd/mm/aaaaQ82-Data Fim:

Q83-Se sim, tipo de Radioterapia:

Q84-Se sim, Dose:

Q85-Quimioterapia: ☐ Sim ☐ Não

T3	C1 Drogas	C2 Dose	C3 Ciclos	C4 Data Início	C5 Data Término
L1					
L2					
L3					
L4					

5/22/2017

>>> Clinical Genomics <<<

L5					
L6					
L7					
L8					
Ln	T3_LnC1	T3_LnC2	T3_LnC3	T3_LnC4	T3_LnC5

OBS: No banco de dados esta tabela acima está codificada da seguinte forma: Tx_LyCz, onde "x", "y" e "z" representam, respectivamente, "x" o número da tabela, "y" o número das novas informações incluídas e "z" o número do campo de informações (drogas,dose,nº dias,ciclos, data início ou data fim). Desta forma, o Ly é a localização de uma nova linha e o Cz é a localização da coluna na tabela. Ex: T3_L3C2 - Tabela3 apresentando informações da linha 3 (y=3) da tabela de quimioterapia da segunda lesão, e da 2ª coluna (C2), que representa o campo com informações de dose.

Q88-Tipo de Quimioterapia (especificar):

G) INFORMAÇÕES SOBRE O ÓBITO

Q87-Óbito: ☐ Sim ☐ Não (preencha os itens abaixo somente se for sim)

Q88-Data Óbito: dd/mm/aaaa

Q89-Tipo de Óbito: ☐ Outra causa ☐ Óbito pela neoplasia

Q90-Se por outra causa, informe qual: