Clinical characteristics and risk fators for fatal and recovery outcomes in Covid-19 patients from Sao Paulo

Amanda L. Pereira[1], **Karla Figueiredo**[2]

[1]Departamento de Engenharia Elétrica
Pontifícia Universidade Católica do Rio de Janeiro
Rio de Janeiro – RJ – Brasil

[2]Departamento de Informática e Ciência da Computação
Instituto de Matemática e Estatística
Universidade do Estado do Rio de Janeiro
Rio de Janeiro – RJ – Brasil

amandalucas@aluno.puc-rio.br, karlafigueiredo@ime.uerj.br

1

Abstract. *In the context of the pandemic caused by the new Coronavirus SARS-CoV-2 (COVID-19), it is relevant to analyze which are the risk factors that most affect patients, along with the groups that are more affected by the disease. In order to apply survival analysis methods to data made available by the Sirio Libanes Hospital (HSL), the Kaplan-Meier estimator and log-rank tests were used for this article. The results indicate a greater probability of survival of female patients, which present a shorter hospitalization time, and a shorter time between hospitalization and death for older patients.*

Resumo. *No contexto da pandemia causada pelo novo Coronavírus SARS-CoV-2 (COVID-19), torna-se relevante analisar quais são os fatores que mais atingem os pacientes, em conjunto com os grupos de risco mais afetados pela doença. Com o objetivo de aplicar métodos de análise de sobrevivência em dados disponibilizados pelo Hospital Sírio Libânes (HSL), foram utilizados para este artigo o estimador de Kaplan-Meier e testes log-rank. Os resultados obtidos indicam uma maior probabilidade de sobrevivência de pacientes do sexo feminino, os quais também apresentam um menor tempo de internação, e um menor intervalo de tempo entre internação e óbito para pacientes de idade avançada.*

1. Introdução

A crise sanitária causada pela pandemia do Coronavírus se espalhou pelo mundo de forma imprescendente, atingindo mais de 200 milhões de pessoas [John Hopkins University 2020]. Apesar da dificuldade inicial da sociedade e dos profissionais de saúde no enfretamento ao vírus, à medida que novos casos foram sendo observados a comunidade científica foi capaz de traçar fatores de risco, que colaboram para um melhor prognóstico e diagnóstico de pacientes [Arruda et al. 2020, Yang et al. 2020].

[1]Cadernos do IME - Série Informática
e-ISSN: 2317-2193 (online)
DOI: 10.12957/cadinf.2021.68556

Com o aumento do número de casos, confirmou-se o impacto da presença de comorbidades e idade avançada na capacidade de recuperação de pacientes contaminados pelo vírus [Martelleto et al. 2021, Prado et al. 2021, Jordan et al. 2020]. Em [Chen et al. 2020], os autores avaliam os dados referente a um grupo de 799 pacientes que ficaram internados em um mesmo hospital em Wuhan, na China, no início da pandemia. Nesse trabalho, observou-se que a idade mediana entre os pacientes que vieram a óbito era significativamente maior que os conseguiram se recuperar da doença. Adicionalmente, notou-se que o sexo masculino representava uma porcentagem maior entre os pacientes que vieram a óbito do que entre os que apresentaram melhora.

Estudos desenvolvidos em outras localidades apresentaram conclusões parecidas, indicando o sexo como um fator relevante no prognóstico do paciente [Olivas-Martínez et al. 2021]. Pesquisas indicam que a interação entre hormônios masculinos e receptores do SARS-CoV-2 tornam homens mais vulneráveis a um quadro complicado da doença [Samuel et al. 2020] e a severidade do quadro do paciente parece estar relacionado a níveis baixos de testorona presentes no organismo [Lanser et al. 2021].

No contexto de buscar auxiliar profissionais da saúde a compreender quais fatores que podem influenciar no desfecho de um paciente que testa positivo para COVID-19, métodos de Análise de Sobrevivência têm sido empregados. Estudos indicaram a influência da idade, complicações renais [Cheng et al. 2020, Di Castelnuovo et al. 2020] e nível de proteína C reativa no sangue [Di Castelnuovo et al. 2020]2. Além destes, outros fatores de risco observados incluem o paciente ser do sexo masculino, apresentar histórico de pneumonia e de hospitalização em serviços públicos de saúde [Salinas-Escudero et al. 2020].

Tendo em vista contribuir para os estudos desenvolvidos nessa área, o presente artigo visa aplicar métodos de Análise de Sobrevivência em dados de pacientes positivos para COVID-19 internados no Hospital Sírio Libanês durante o ano de 2020, buscando analisar variáveis relacionadas ao prognóstico dos mesmos, evidenciado pela evolução a óbito ou a melhora. Nas análises realizadas, as variáveis consideradas foram a idade e o sexo do paciente.

O restante do trabalho está distribuído em mais quatro seções: na seção 2 é apresentada uma breve introdução dos fundamentos técnicos necessários para melhor compreensão dos métodos e modelos desenvolvidos neste trabalho. A metodologia utilizada para solução do problema proposto, assim como sua aplicabilidade à esfera do problema é descrita na terceira seção. Os estudos de casos são apresentados e os resultados são discutidos na seção 4. Por fim, a última seção encerra o trabalho apresentando as conclusões e perspectivas de novos trabalhos.

2. Metodologia

2.1. Análise de Sobrevivência

Análise de Sobrevivência, também chamada de Análise de Sobrevida, consiste em um conjunto de métodos estatísticos para análise de dados com o objetivo de extrair a informação de tempo até que um evento de interesse ocorra [Kleinbaum and Klein 2010]. O evento de interesse compreende qualquer evento que possa ocorrer a um indivíduo incluído na análise, sendo que neste estudo considera-se os eventos melhora e óbito dos

pacientes analisados. Para realização das análises foi utilizado o pacote scikit-survival [Pölsterl 2020] 0.15.0 em um Ambiente Anaconda com Python 3.9.4.

2.1.1. Estimador de Kaplan-Meier

O método conhecido como estimador de Kaplan-Meier consiste em uma estimação não-paramétrica da função de sobrevivência a partir de dados de vida útil. Nesse contexto, a função de sobrevivência busca estimar a fracção de pacientes que vivem durante um determinado período de tempo, chamado de tempo de sobrevida. A estimação realizada por esse método se baseia na consideração de que o tempo de sobrevida até cada unidade de tempo t é independente da sobrevivência até as outras unidades de tempo anteriores, ti [Kaplan and Meier 1958].

A probabilidade de um indivíduo sofrer o evento, que no caso deste trabalho é o evento de óbito ou de melhora, em um tempo t é o produto da probabilidade de se chegar até cada um dos tempos anteriores t_i. O estimador da distribuição $S_{KM}(t)$ é o produto das probabilidades de sobrevivência a cada tempo $t_i < t$, onde t_i são os tempos de observação da ocorrência dos eventos [Shimakura et al. 2005]. Considerando que $R(t_i)$ é o total de indvidíduos a risco de sofrer ocorrência do evento no tempo t_i, tem-se:

$$S_{KM}(t) = \prod_{t_i \leq t} \frac{R(t_i) - N(t_i)}{R(t_i)} \tag{1}$$

2.2. Testes Log-rank

No processo de análise da sobrevida deve ser verificado se há diferença estatísticas entre as curvas de Kaplan-Meier, onde cada curva representa uma população, ou seja, grupos distintos. Nesse caso o teste Log-rank pode ser usado para avaliar se duas ou mais curvas são estatisticamente diferentes [Fleming and Harrington 1981]. Ao avaliar as curvas, afirmar que estas são estatisticamente diferentes significa indicar, através da aplicação do teste de Log-rank, que as populações são diferentes.

2.3. Base de Dados

O conjunto de dados escolhido foi o Dados COVID Hospital Sírio-Libanês [Sírio Libanês 2020], disponibilizada no repositório COVID-19 DataSharing/BR, uma iniciativa entre a Fundação de Pesquisa de São Paulo (FAPESP) e a Universidade de São Paulo (USP) [Mello et al. 2020]. A base consiste em três arquivos contendo os dados e um dicionário de dados, que indica como relacionar as informações contidas em cada arquivo.

A primeira tabela, chamada ”Pacientes”, contém as seguintes informações: um número identificador do paciente (anonimizado nessa e nas demais tabelas), sexo, ano de nascimento, país de residência, unidade da federação de residência do paciente, cidade e CEP parcial. A tabela ”Exames”contém em seus registros o ID do paciente, ID de atendimento, data da coleta, local de coleta, descrição do exame realizado, descrição do analito referente ao exame realizado, resultado do exame, unidade de medida do analito e valores de referência.

A terceira e última tabela da base de dados, chamada ”Desfechos”, traz as informações de ID de atendimento (correlacionada com a tabela ”Exames”), data de realização do atendimento, tipo de atendimento, identificação e descrição da clínica onde aconteceu o atendimento, e a data e descrição do desfecho. Com a informação de ID do paciente, é possível correlacionar as três tabelas disponibilizadas pela base de dados.

2.4. Pré-processamento dos Dados

A primeira etapa do estudo foi a de pré-processamento, onde primeiramente se definiu os atributos de interesse de cada tabela. Nessa etapa, foram removidos da análise os atributos referentes à origem e residência do paciente da tabela ”Pacientes”, o atributo referente ao local de coleta do exame da tabela ”Exames”e as colunas referentes à clínica de atendimento do paciente na coluna ”Desfechos”.

A etapa seguinte consistiu em descartar algumas amostras do conjunto de dados. Essas foram excluídas por conterem informações anonimizadas de ano de nascimento (”YYYY”na tabela de Pacientes). Na tabela ”Desfechos”, a data relativa ao desfecho da amostra contém o valor ”DDMMAA”para as linhas relativas à pacientes que vieram a óbito. Ou seja, não é possível, a partir dessa coluna, determinar a data em que o paciente veio a óbito. Nesses casos, foi mantida a string original no valor do atributo, e considerou-se a data do óbito referente à data da última coleta de exames de cada paciente.

Em seguida, as colunas referentes a medidas de tempo decorrido (data de atendimento, data da coleta, ano de nascimento) foram discretizadas: as datas foram convertidas para um valor inteiro, referente ao número de dias decorridos a partir do dia 01/11/2019, que marca o início da coleta dos dados da base. As informações contidas no atributo ”ano de nascimento”foram discretizadas, separando-se os dados em seis grupos de acordo com a faixa etária: grupo 0 (nascidos até o ano de 1930), grupo 1 (nascidos de 1931 a 1950), grupo 2 (1951 a 1970), grupo 3 (1971 a 1990), grupo 4 (1991 a 2010) e grupo 5 (nascidos a partir de 2011), conforme indicado na tabela 1.

Tabela 1. Agrupamento por Faixa Etária

Idade (anos)	Grupo
Acima de 90	0
71 a 90	1
51 a 70	2
31 a 50	3
11 a 30	4
Abaixo de 10	5

Na tabela ”Exames”, foram filtrados apenas os exames relativos à COVID-19. A partir de uma transformação aplicada ao atributo referente a esses exames, criou-se uma nova coluna com os resultados discretizados do(s) exame(s) de COVID-19 de cada paciente: 1 para amostras reagentes para o vírus, e 0 para as amostras negativas.

Após a limpeza e discretização dos atributos, o próximo passo foi gerar as variáveis necessárias para a realização da análise de sobrevivência. Para os eventos consi-

derados nesse trabalho - óbito e melhora do paciente, separadamente -, foram necessários criar dois atributos para cada: um atributo binário indicando se o evento ocorreu ou não para dado paciente, e um atributo indicando o tempo ocorrido até o evento caso este tenha ocorrido.

Para a análise dos pacientes que vieram à óbito, é necessário realizar a censura dos pacientes que receberam alta. Logo, estes foram removidos desta análise. De modo análogo, ao analisar os pacientes que apresentaram melhora, desconsiderou-se os pacientes que vieram a óbito.

3. Resultados

3.1. Curvas de Kaplan-Meier

As Figuras 1 e 2 apresentam as curvas obtidas a partir do estimador de Kaplan-Meier para o evento de óbito. A escala de tempo (eixo x) para todos os gráficos apresentados é em dias, e o número n na legenda indica o número de pacientes pertencente a cada grupo. Após a censura dos dados de pacientes que vieram a receber alta, a base utilizada para essa análise contou com apenas 103 pacientes, dos quais 72 vieram a óbito no mesmo dia de admissão no hospital (nesse caso, o tempo até o evento é 0).

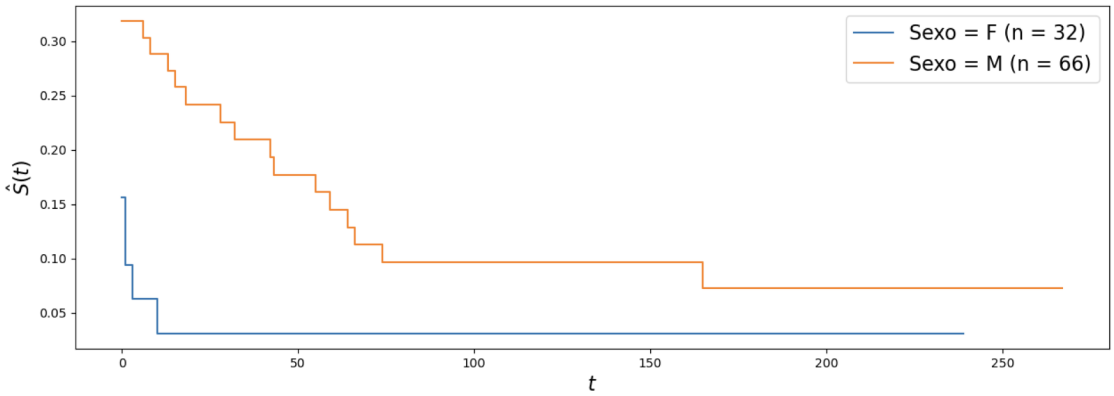

Figura 1. Curva de Sobrevivência de Kaplan-Meier para o evento Óbito, com grupos separados por sexo.

Para os pacientes considerados na análise, nota-se que a probabilidade de sobrevivência estimada para os do sexo feminino se estabiliza em um valor superior aos do sexo masculino a partir dos 80 dias de hospitalização. Devido ao grande número de pacientes com $t = 0$, a curva estimada já tem seu $\hat{S(t)}$ iniciando em valores baixos, com $\hat{S(0)} = 0.31$ para mulheres e $\hat{S(0)} = 0.15$ para homens (Figura 1).

Agrupando os pacientes por idade (Figura 2), as curvas de cada grupo indicam uma tendência de que quanto maior a faixa etária do indivíduo, menor o tempo deste até o óbito. Não está plotada a curva referente a indivíduos pertencentes ao grupo 4 ou 5 (pacientes nascidos a partir de 1991) por não haver nenhum paciente que se encaixou nesse grupo. Ou seja, vieram a óbito apenas pacientes acima de 30 anos entre os observados.

As figuras 3 e 4 apresentam as curvas de Kaplan-Meier para o evento de melhora, que compreende todos os tipos de alta presentes na base dados. Para essa análise foram

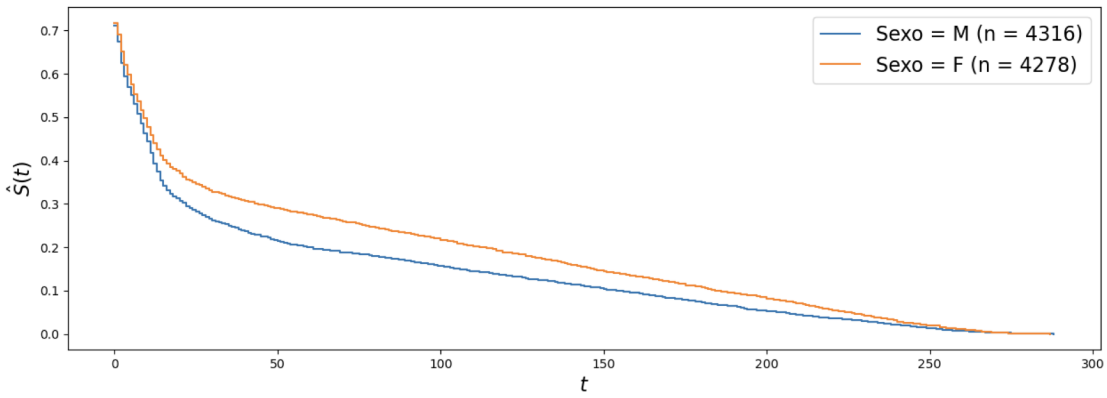

Figura 2. Curva de Sobrevivência de Kaplan-Meier para o evento Óbito, com grupos separados por idade.

censurados todos os pacientes que vieram à óbito, totalizando 8611 pacientes. A curva gerada para a tabela agrupada por sexo indica que, à medida que o tempo de hospitalização de um paciente no hospital aumenta, há maior probabilidade do mesmo apresentar melhora se este for do sexo feminino.

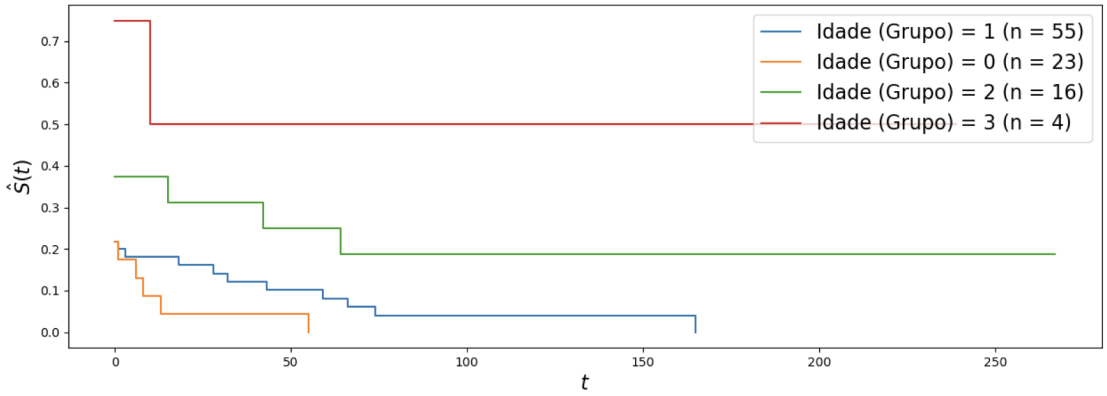

Figura 3. Curva de Sobrevivência de Kaplan-Meier para o evento Melhora, com grupos separados por sexo.

A curva gerada para o evento de melhora com grupos separados por idade indicam um tempo menor até a melhora para pacientes pertencentes do grupo 5, isto é, nascidos a partir de 2011. As curvas estimadas para os grupos 1, 2, 3 e 4 apresentam comportamento mais suave devido ao maior número de pacientes observados pertencentes a essas faixas etárias, diferente da curva estimada para o grupo 0 - referente a pacientes nascidos até 1930, com apenas 46 pacientes. Esse gráfico também indica que, pacientes dos grupos intermediários, nascidos entre 1931 a 2010, apresentam um tempo mais longo até o evento de melhora do que as demais faixas etárias, se estendendo até aproximadamente 270 dias.

3.2. Testes Log-rank

O teste de log-rank tem como hipótese nula que as taxas de risco em todos os grupos são iguais. Os resultados obtidos para ambos eventos considerados estão apresentados

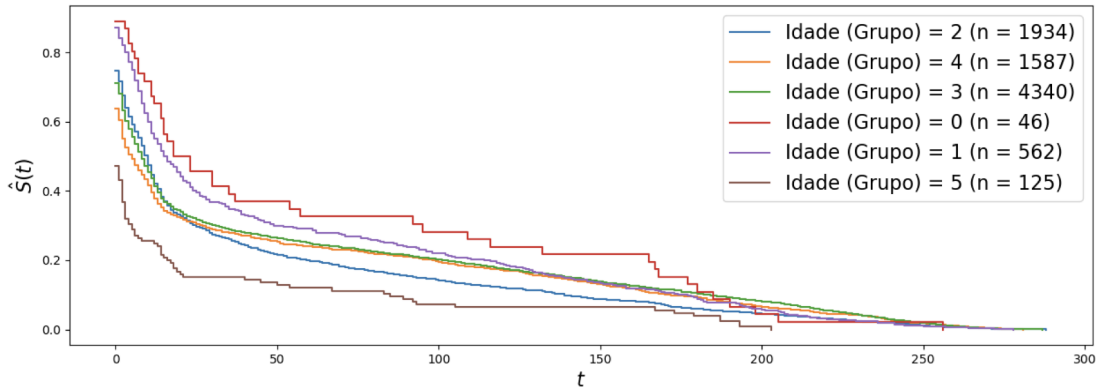

Figura 4. Curva de Sobrevivência de Kaplan-Meier para o evento Melhora, com grupos separados por idade.

na Tabela 2. Em relação ao evento de óbito, o teste de log-rank retornou um P value de 0.022 para o sexo, e de 0.0052 para o atributo referente aos grupos de idade, ou seja, o evento apresenta significância estatística para estes dois atributos ao se avaliar o tempo de sobrevivência de um indivíduo hospitalizado com COVID-19.

Aplicando o mesmo teste nos dados e considerando o evento de melhora do paciente, os valores obtidos foram de 4.14×10^{-11} e de 1.55×10^{-11} para o atributo referente a sexo e a idade, respectivamente. Esses valores indicam uma reafirmação da diferença observada entre as curvas para cada grupo nos gráficos de Kaplan-Meier, e negam a hipótese nula de que as taxas de risco para todos os grupos são iguais [Bewick et al. 2004].

Tabela 2. Testes de Log-rank

Evento	Atributo	P
Óbito	Sexo	0.022
	Idade (grupos)	0.0052
Melhora	Sexo	4.14×10^{-11}
	Idade (grupos)	1.55×10^{-11}

4. Conclusões e Trabalhos Futuros

Os resultados apresentados indicam que as variáveis analisadas - idade e sexo - estão correlacionados com o tempo de hospitalização de cada paciente, e também com o seu desfecho final, seja este positivo ou negativo. As análises realizadas incluíram apenas a utilização da informação relativa a exames de COVID-19 contidas na base.

O próximo passo do trabalho seria incluir os atributos relativos a mais exames ou aplicar a *pipeline* a um conjunto de dados que contenha informações sobre comorbidades, de forma que possibilite a extração de conhecimento relativo à influência desses fatores nos desfechos considerados. Por exemplo, ao se analisar a influência do sexo no desfecho do indivíduo, os dados não contavam com informações referentes à fatores de risco como tabagismo, que por ser um hábito mais observável em homens do que mulheres [Ritchie 2019] poderia indicar uma razão pela qual o público feminino apresenta maior probabilidade de sobrevivência.

Referências

Arruda, D. É. G., Martins, D. D. S., da Silva, I. F. M., and de Sousa, M. N. A. (2020). Prognóstico de pacientes com covid-19 e doenças crônicas: uma revisão sistemática. *Comunicação em Ciências da Saúde*, 31(03):79–88.

Bewick, V., Cheek, L., and Ball, J. (2004). Statistics review 12: survival analysis. *Critical care*, 8(5):1–6.

Chen, T., Wu, D., Chen, H., Yan, W., Yang, D., Chen, G., Ma, K., Xu, D., Yu, H., Wang, H., Wang, T., Guo, W., Chen, J., Ding, C., Zhang, X., Huang, J., Han, M., Li, S., Luo, X., Zhao, J., and Ning, Q. (2020). Clinical characteristics of 113 deceased patients with coronavirus disease 2019: retrospective study. *BMJ*, 368.

Cheng, Y., Luo, R., Wang, K., Zhang, M., Wang, Z., Dong, L., Li, J., Yao, Y., Ge, S., and Xu, G. (2020). Kidney disease is associated with in-hospital death of patients with covid-19. *Kidney international*, 97(5):829–838.

Di Castelnuovo, A., Bonaccio, M., Costanzo, S., Gialluisi, A., Antinori, A., Berselli, N., Blandi, L., Bruno, R., Cauda, R., Guaraldi, G., et al. (2020). Common cardiovascular risk factors and in-hospital mortality in 3,894 patients with covid-19: survival analysis and machine learning-based findings from the multicentre italian corist study. *Nutrition, Metabolism and Cardiovascular Diseases*, 30(11):1899–1913.

Fleming, T. R. and Harrington, D. P. (1981). A class of hypothesis tests for one and two sample censored survival data. *Communications in Statistics-Theory and Methods*, 10(8):763–794.

John Hopkins University (2020). Covid-19 dashboard. Disponível em `https://coronavirus.jhu.edu/map.html`.

Jordan, R. E., Adab, P., and Cheng, K. K. (2020). Covid-19: risk factors for severe disease and death. *BMJ*, 368.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282):457–481.

Kleinbaum, D. G. and Klein, M. (2010). *Survival analysis*. Springer.

Lanser, L., Burkert, F. R., Thommes, L., Egger, A., Hoermann, G., Kaser, S., Pinggera, G. M., Anliker, M., Griesmacher, A., Weiss, G., et al. (2021). Testosterone deficiency is a risk factor for severe covid-19. *Frontiers in Endocrinology*, 12:731.

Martelleto, G. K. S., Alberti, C. G., Bonow, N. E., Giacomini, G. M., Neves, J. K., de Miranda, E. C. A., da Silveira, I. D., and de Macedo, I. C. (2021). Principais fatores de risco apresentados por pacientes obesos acometidos de covid-19: uma breve revisão. *Brazilian Journal of Development*, 7(2):13438–13458.

Mello, L. E., Suman, A., Medeiros, C. B., Prado, C. A., Rizzatti, E. G., Nunes, F. L. S., Barnabé, G. F., Ferreira, J. E., Sá, J., Reis, L. F. L., Rizzo, L. V., Sarno, L., de Lamonica, R., Maciel, R. M. d. B., Cesar-Jr, R. M., and Carvalho, R. (2020). Opening Brazilian COVID-19 patient data to support world research on pandemics.

Olivas-Martínez, A., Cárdenas-Fragoso, J. L., Jiménez, J. V., Lozano-Cruz, O. A., Ortiz-Brizuela, E., Tovar-Méndez, V. H., Medrano-Borromeo, C., Martínez-Valenzuela, A.,

Román-Montes, C. M., Martínez-Guerra, B., et al. (2021). In-hospital mortality from severe covid-19 in a tertiary care center in mexico city; causes of death, risk factors and the impact of hospital saturation. *Plos one*, 16(2):e0245772.

Pölsterl, S. (2020). scikit-survival: A library for time-to-event analysis built on top of scikit-learn. *Journal of Machine Learning Research*, 21(212):1–6.

Prado, P. R. d., Gimenes, F. R. E., Lima, M. V. M. d., Prado, V. B. d., Soares, C. P., and Amaral, T. L. M. (2021). Fatores de risco para óbito por covid-19 no acre, 2020: coorte retrospectiva. *Epidemiologia e Serviços de Saúde*, 30.

Ritchie, H. (2019). Who smokes more, men or women? Disponível em `https://ourworldindata.org/who-smokes-more-men-or-women`.

Salinas-Escudero, G., Carrillo-Vega, M. F., Granados-García, V., Martínez-Valverde, S., Toledano-Toledano, F., and Garduño-Espinosa, J. (2020). A survival analysis of covid-19 in the mexican population. *BMC Public Health*, 20(1):1–8.

Samuel, R. M., Majd, H., Richter, M. N., Ghazizadeh, Z., Zekavat, S. M., Navickas, A., Ramirez, J. T., Asgharian, H., Simoneau, C. R., Bonser, L. R., et al. (2020). Androgen signaling regulates sars-cov-2 receptor levels and is associated with severe covid-19 symptoms in men. *Cell Stem Cell*, 27(6):876–889.

Shimakura, S. E., Carvalho, M., Andreozzi, V., Codeço, C., and Barbosa, M. (2005). Análise de sobrevida: Teoria e aplicações em saúde. *Rio de Janeiro: Editora Fiocruz*.

Sírio Libanês, H. (2020). Dados covid hospital sírio-libanês. Dados disponíveis em `https://repositoriodatasharingfapesp.uspdigital.usp.br/handle/item/97`.

Yang, J., Zheng, Y., Gou, X., Pu, K., Chen, Z., Guo, Q., Ji, R., Wang, H., Wang, Y., and Zhou, Y. (2020). Prevalence of comorbidities in the novel wuhan coronavirus (covid-19) infection: a systematic review and meta-analysis. *Int J Infect Dis*, 10(10.1016).