Análise do uso de quimioterápicos no SUS do RJ entre 2013 e 2022

Luiz Gabriel da S. Samuel[1], Rafaela P. M. Fernandes[2], Tiago R. de Matos[3]

[1]Instituto de Computação – Universidade Federal Fluminense (UFF)
Caixa Postal 100.175 – 24.210-346 – Niterói – RJ – Brazil

{luizgss, rafaelapecanha, ti_matos}@id.uff.br

Abstract. *This article weaves the creation of an Artificial Intelligence model through a logistic regression with supervised learning based on a processed dataset cancer, created based on DataSUS, aiming at creating a model capable of predicting the planning of the patient based on data such as ethnicity, initial procedure, initial staging and age.*

Resumo. *Este artigo tece a criação de um modelo de Inteligência Artificial por meio de uma regressão logística com aprendizado supervisionado com base em um dataset relacionado ao câncer de mama, criado com base no DataSUS, visando a criação de um modelo capaz de predizer o prognóstico do paciente com base em dados como etnia, procedimento inicial, estadiamento de início e idade.*

1. Introdução

Segundo relatório anual do INCA, em 2022, projeta-se 66.280 novos casos, para uma taxa de incidência ajustada de 43,74 por 100.000 mulheres (INCA, 2019a). As taxas brutas de incidência e os números estimados de novos casos são importantes para estimar a gravidade da doença na região e para desenvolver planos de ação locais.

Desta forma, este artigo relata a elaboração e criação de um modelo de Regressão Logística com base em um dataset criado pelo Apac de Quimioterapia, obtidos através do DataSUS entre 2013 e 2022, relacionado ao câncer de mama e todos os dados que foram guardados desses pacientes sem sua identificação. O objetivo principal deste artigo e projeto era a criação de um modelo de IA que pudesse ser utilizado para indicar qual seria o prognóstico do paciente com base em informações passadas pelo paciente. O problema envolve regressão logística com aprendizado supervisionado, onde uma amostra de pacientes contém informações sobre características que classificam as amostras de acordo com seus resultados prováveis. Ou seja, dada uma instância e seus atributos preditores, preveja os atributos de destino correspondentes.

O modelo desenvolvido pode ser encontrado com o seu código de treinamento e seu dataset manipulado no github do projeto. Como pode ser observado, este é um projeto

ainda embrionário, porém com grande espaço de crescimento, tanto em código, quanto em manipulação de dataset. O desenvolvimento das atividades se deu por meio da linguagem de programação Python, utilizando os recursos da biblioteca de aprendizado de máquina scikit-learn, via Jupyter Notebook.

2. Revisão de Literatura

2.1 Metodologia

A base de dados utilizada foi a PubMed, no qual os descritores escolhidos foram “Artificial Intelligence”, "Diagnosis" e "Logistic Models", o operador booleano utilizado foi “AND”. Os registros identificados por meio de pesquisa de banco de dados foram n=161. Logo, o critério para exclusão foi livros e documentos, ensaios clínicos, ensaios clínicos controlados e randomizados, já o critério para inclusão foi revisão e revisão sistemática publicadas em 5 anos, sendo encontrados apenas 2 estudos.

2.2 Resultados

A estratificação de risco é um sistema pelo qual separações clinicamente significativas de riscos podem ser alcançadas em um grupo de indivíduos semelhantes. Apesar do domínio da regressão logística paramétrica na previsão de risco, o uso de métodos não paramétricos e semi paramétricos, incluindo redes neurais artificiais, está aumentando. Esses métodos de aprendizado estatístico e aprendizado de máquina, juntamente com regras simples, são chamados coletivamente de "inteligência artificial" (IA). A IA requer conhecimento da validade do estudo, compreensão das métricas do modelo e determinação se e até que ponto um modelo pode e deve ser aplicado ao paciente ou população em consideração. Mais investigação é necessária, especialmente em termos de validação de modelos e avaliação de impacto.

Ademais, ambos os modelos de logística e inteligência artificial podem ser utilizados de forma eficaz para distinguir diagnóstico e prognóstico em maligno ou benigno. Ambos os métodos são adequados para analisar a quantidade de dados excede em muito a mente humana. Acontece que ambos os métodos são capazes de fornecer classificação automática de alta precisão, como tal, espera-se que desempenhem um papel importante em diagnósticos futuros. Essas tecnologias nunca devem ser consideradas um substituto, mas sim uma ferramenta complementar para melhorar a prática clínica.

De acordo com Tu et al, existem várias vantagens sobre os modelos logísticos. Sendo essas, exige menos treinamento estatístico formal, menos capacidade de detectar explicitamente relacionamentos não lineares complexos entre variáveis dependentes e independentes, e a possibilidade para detectar todas as interações entre as variáveis.

A regressão logística demonstra vantagens quando o problema em questão é a de predizer um prognóstico ou diagnóstico. No entanto, mesmo com bom desempenhos devido a alta precisão na classificação dos pacientes. A tecnologia ainda é indicada a ser utilizada apenas como ferramenta auxiliar para o profissional da saúde.

3. Metodologia da Análise Experimental

O conjunto de dados do Apac de Quimioterapia do Datasus consiste de amostras de pacientes (instâncias) contendo informações acerca de algumas das suas características (atributos preditores) e, também, seu respectivo desfecho (atributo alvo/classe): melhora, alta, piora e óbito, no qual criamos para que fosse possível realizar a predição.

O problema envolve uma regressão logística com aprendizado supervisionado, no qual constitui-se em uma amostra de pacientes contendo as informações sobre as características, classificando a amostra de acordo com os possíveis desfechos dos pacientes. Ou seja, dada uma instância e seus atributos preditores, predizer o atributo alvo correspondente.

3.1. Preparação dos Dados

Reunimos todas as bases e as limpamos, para deixar o dataset no github, o deixamos em formato pickle, e a partir dessa base, começamos a preparar nossas informações para a análise e processamento de dados, verificando quais seriam úteis e quais precisamos pré-processar.

O conjunto de dados em questão consiste em 273107 instâncias compostas, cada uma, por 64 atributos. Na preparação dos dados foi realizada a análise exploratória na qual verificou-se que dos 64, muitos foram retirados por serem dados administrativos que não contribuem com informações relevantes para a determinação do prognóstico do paciente, restando 46 atributos. Também foi excluído o campo “codigo_idade”, para que não houvesse redundância com o campo “numero_idade”, significando a idade especificamente. Ainda levando em consideração o campo idade foram retirados os pacientes acima de 100 anos de idade. Como o assunto abordado é câncer de mama e a incidência em pacientes do sexo masculino é baixa, então foi decido retirar os pacientes do sexo masculino para que não houvesse ruído e algum tipo de enviesamento das análises.

Outrossim, o atributo “esquema_terapeutico” foi retirado por que era aberto com valores inconsistentes. O atributo “municipio_residencia” foi retirado, considerando apenas o campo “trata_fora_municipio” a fim de apenas verificar se o paciente está realizando tratamento dentro ou fora do próprio município de residência. Os campos “dif_estadiamento”, “indicador_obito”, “indicador_alta” e “indicador_encerramento” serviram para criar o atributo alvo denominado desfecho que possui as classes “piora”, “obito”, “alta” e “melhora” respectivamente. Com isso, foram observados 211636 registros de pacientes em que foi classificado como “SemDesfecho”, que foram retirados do dataset de análise. Logo, os atributos “cns_paciente”, “numero_idade”, “raca_cor”, “trata_fora_municipio”, “estadiamento_inicio”, que são os atributos preditivos e numéricos, o “procedimento_inicial” que era categórico e foi “dummizado” e, por fim, o desfecho, que é o atributo alvo e categórico, foram os escolhidos para o processamento dos dados. Ademais, foi realizada uma agregação, visto que o CNS (Cadastro Nacional de Saúde) aparece inúmeras vezes no dataset, ou seja, estava sendo contabilizada a mesma pessoa mais de uma vez. Desta forma, após a agregação ficaram

5.621 instâncias com 3423 ocorrências da classe Melhora (60.90%), com 1815 ocorrências da classe Piora (32.29%), com 195 ocorrências da classe Alta (3.57%) e também com 188 ocorrências da classe Óbito (3.34%), isto é, o conjunto de dados é desbalanceado e possui um erro majoritário de 39.10%.

3.2 Processamento de Dados

Antes da etapa de construção de modelos, é crucial observar que: pela natureza do problema em questão, o mais interessante é que o modelo construído seja capaz de gerar uma quantidade mínima de falsos negativos.Visto que, o esperado é que não hajam prognósticos errados, a fim de que não haja alteração das intervenções da saúde do paciente.

Também é necessário levar em consideração a divisão entre dados utilizados para teste do modelo gerado (20%) e os dados utilizados na criação do modelo(80%). Com isso foi aplicado a função logística, a fim de termos um modelo com regressão logística, sendo gerada uma matriz de confusão.

4. Resultados Obtidos

Após analisarmos os dados e prepararmos os modelos, conseguimos chegar a conclusão de um modelo de imagem otimista com relação a sua capacidade de análise.

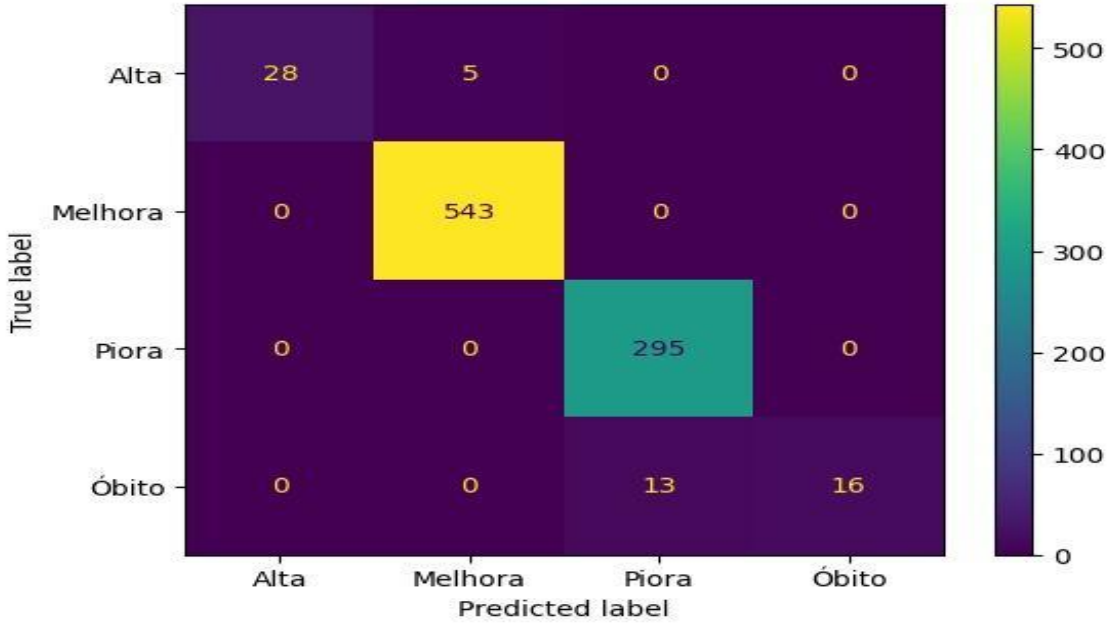

Figura 1. Matriz de confusão

Uma matriz de confusão é uma tabela que indica os erros e acertos do seu modelo, comparando com o resultado esperado (ou etiquetas/*labels*). Com isso, a Figura 1 demonstra a matriz de confusão gerada por meio dos dados do dataset escolhido. É possível observar pela matriz de confusão que tivemos alguns falsos positivos e falsos

negativos, sendo assim a mesma permitiu por meio de suas informações gerar métricas de avaliação como acurácia, precisão, recall e f1-score.

Neste caso, foi possível observar acurácia de 98%, com precisão de 98%, como é possível observar na Tabela 1. Um ponto importante a ser observado é que no nosso dataset em questão em que os Falsos Negativos são considerados mais prejudiciais que os Falsos Positivos, ou seja, o modelo deve encontrar todos os pacientes doentes, mesmo que classifique alguns saudável como doentes (situação de Falso Positivo) no processo. Isto é, o modelo deve ter alto recall, pois classificar pacientes doentes como saudáveis pode ser uma tragédia. A métrica *recall* segundo a Tabela 1 tem seus valores majoritariamente altos, no entanto, apenas na classe "Óbito" obteve 55%, o que é ruim porque indica que não está classificando adequadamente o paciente, visto que está tendo uma subvalorização de "Obito", o que está gerando uma supervalorização de *Piora*.

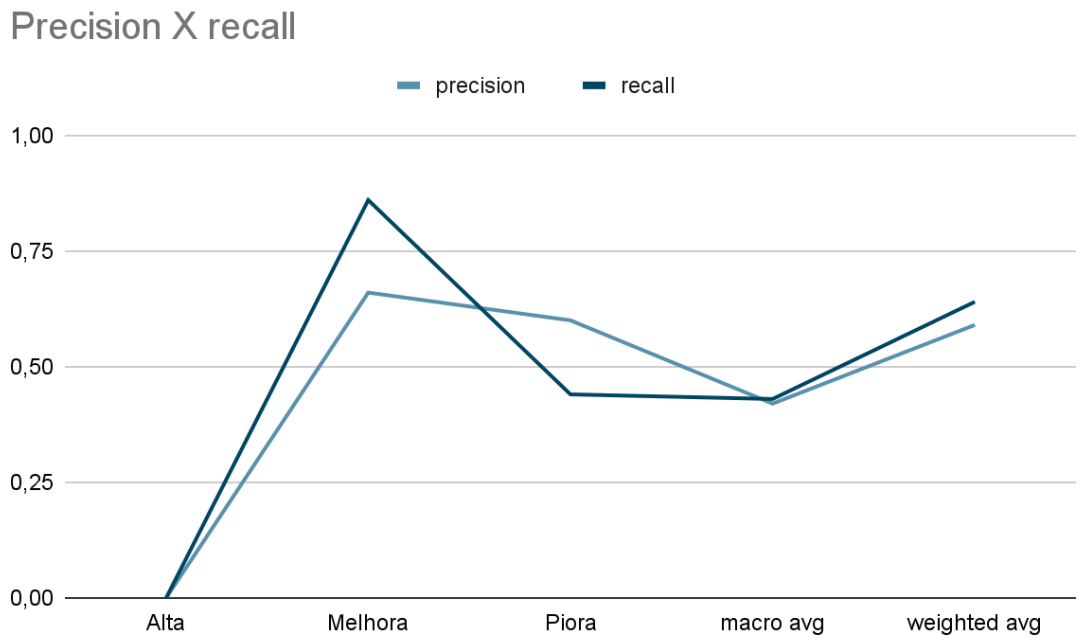

Figura 2. Gráfico comparando Precision e Recall

	Precision	recall	f1-score	support
Alta	1.00	0.85	0.92	33
Melhora	0.99	1.00	1.00	543
Piora	0.96	1.00	0.98	295
Obito	1.00	0.55	0.71	29
Accurary			0.98	900
Macro avg	0.99	0.85	0.90	900
Weighted avg	0.98	0.98	0.98	900

Tabela 1. Métricas de Avaliação do dataset

Como o dataset de escolha estava desbalanceado tentamos tratar o desbalanceamento aplicando *downsampling*, que é a técnica de redução da taxa de amostragem. Isso é feito simplesmente separando uma amostra a cada N, ou seja, foram removidos dados da classe majoritária. No entanto, após aplicar *downsampling* foi gerada uma outra matriz de confusão com base nos dados balanceada que é possível observar na Figura 3.

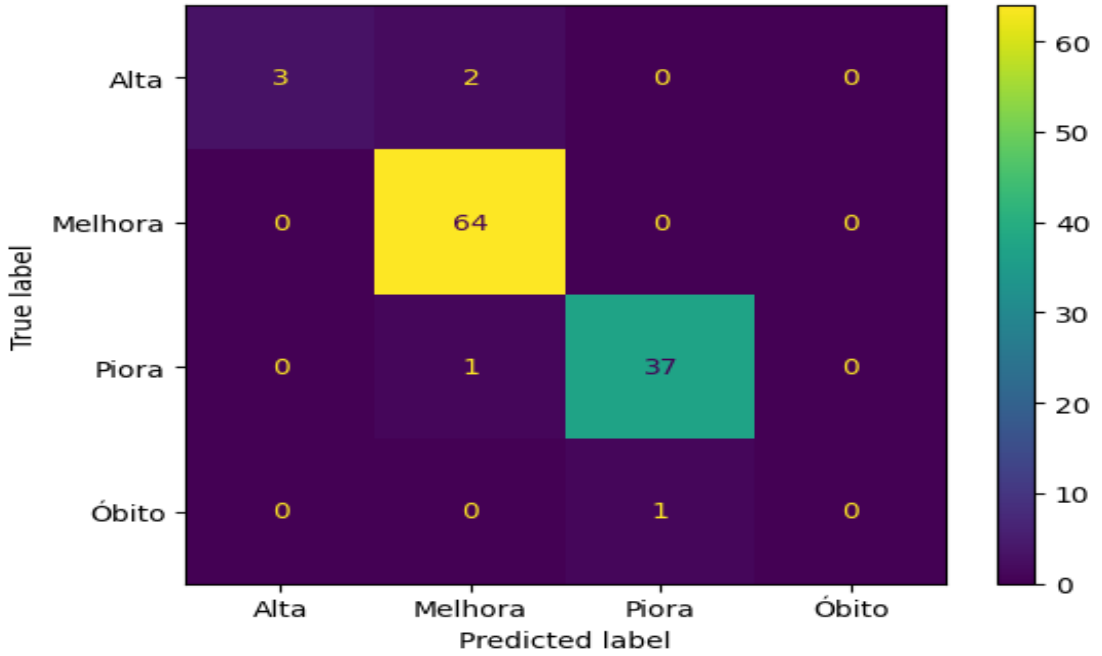

Figura 3. Matriz de confusão após *downsampling*

Com a nova matriz foi possível gerar métricas de avaliação, no qual foi obtido 96% de acurácia, o que é um valor próximo a acurácia já obtida antes do balanceamento do dataset. No entanto, foi obtido também precisão de 73%, o que é uma mudança

significativa quando comparado o valor de 98% obtido anteriormente. Com o balanceamento é possível observa na matriz de confusão que nenhum caso foi classificado como “Óbito”, ou seja, houve uma supervalorizaçãoem outra classe.

Na Figura 4 é possível observar a matriz de correlação, a mesma mostra os valores de correlação de Pearson, que medem o grau de relação linear entre cada par de variáveis. Os valores de correlação podem cair entre -1 e +1. Se as duas variáveis tendem a aumentar e diminuir juntas, o valor de correlação é positivo. Se uma variável aumenta enquanto a outra variável diminui, o valor de correlação é negativo. Sendo possível observar, por exemplo, que existe uma relação linear negativa para os seguintes pares, com coeficientes negativos de correlação de Pearson, os códigos são relacionados aos respectivos procedimentos como consta no dicionário encontrado no repositório, sendo os mais relevantes destacar procediemento_inicial_0304040029 (QUIMIOTERAPIA DO CARCINOMA DE MAMA (PRÉVIA)) ,procedimento_inicial_0304020338 ('HORMONIOTERAPIA DO CARCINOMA DE MAMA AVANÇADO - 2ª LINHA) e procediemento_inicial_0304020133 (QUIMIOTERAPIA DO CARCINOMA DE MAMA AVANÇADO -1ª LINHA) que possuem diversos coeficientes negativos com outros procedimentos iniciais.

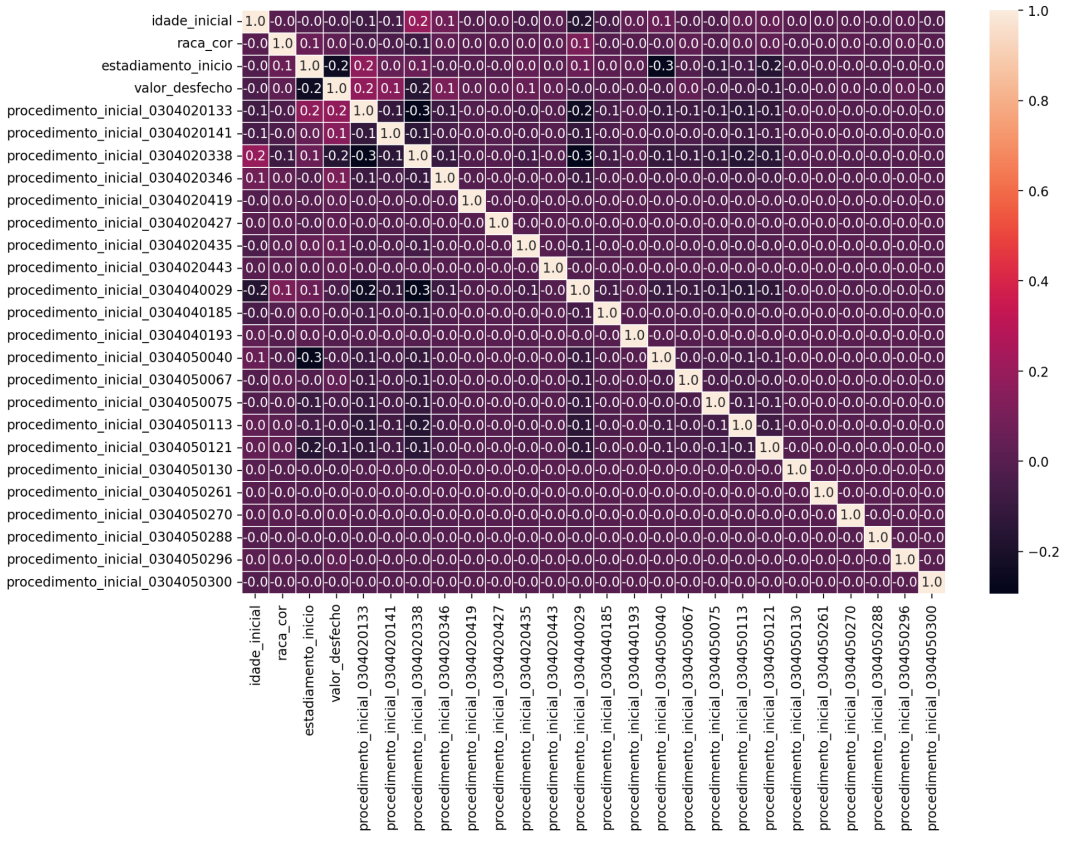

Figura 4. Matriz de correlação

5. Conclusão

Portanto, com base no modelo desenvolvido podemos identificar uma forte inter-relação entre idade e tipo de procedimento realizado, bem como a etnia não teve muita interferência nos procedimentos iniciais.

Outro fato é a necessidade de sanitização da base, tendo em vista que mesmo uma base extremamente diversa como do Datasus detém campos "livres", o que cria uma certa barreira para um modelo inicial e sua posterior análise. Um exemplo desses campos é o campo do esquema terapêutico.

Em síntese, o DataSUS disponibiliza informações que podem servir para subsidiar análises objetivas da situação sanitária, tomadas de decisão baseadas em evidências e elaboração de programas de ações de saúde. Sendo assim, o SUS atualmente fornece inúmeros dados que permitem que diversos trabalhos de análise possam ser realizados acerca destes dados, para a construção de ferramentas de auxílio de prognóstico e diagnóstico.

Portanto, um possível trabalho futuro seria o uso do Datasus para a criação de um modelo capaz de auxiliar na predição de um planejamento de medicação para o paciente, tendo em vista que a vastidão da base também engloba informações como os medicamentos utilizados pelo paciente e em relação ao procedimentos realizados.

6. Referências

Repositório com dados do SUS do RJ entre 2013 e 2022. GitHub, 2023. Disponível em: https://github.com/TiamatCod3/ia-saude/treeqmaster/dados

Transferência de Arquivos. DATASUS, 2023. Disponível em: https://datasus.saude.gov.br/tranferencia-de-arquivos/#. Acesso em: 20 de junho de 2023

Repositório com Análise de dados do SUS do RJ entre 2013 e 2022. GitHub, 2023. Disponível em: https://github.com/TiamatCod3/ia-saude.git

Instituto Nacional de Câncer. DADOS E NÙMEROS SOBRE CÂNCER DE MAMA: Relatório anual 2022. Brasil, 2023.

Song X, Liu X, Liu F, Wang C. Comparison of machine learning and logistic regression models in predicting acute kidney injury: A systematic review and meta-analysis. Int J Med Inform. 2021 Jul;151:104484. doi: 10.1016/j.ijmedinf.2021.104484. Epub 2021 May 8. PMID: 33991886.

Grigore M, Popovici RM, Gafitanu D, Himiniuc L, Murarasu M, Micu R. Logistic models and artificial intelligence in the sonographic assessment of adnexal masses - a systematic review of the literature. Med Ultrason. 2020 Nov 18;22(4):469-475. doi: 10.11152/mu-2538. Epub 2020 Jun 29. PMID: 32905566.