Universidade de São Paulo

Faculdade de Saúde Pública

Programa de Pós-Graduação em Saúde Pública

Inteligência Artificial para a vigilância de doenças crônicas não-transmissíveis

Gabriel Ferreira dos Santos Silva

Tese apresentada ao Programa de Pós-Graduação em Saúde Pública da Faculdade de Saúde Pública da Universidade de São Paulo para obtenção do título de Doutor em Ciências.

Área de concentração: Saúde Pública.

Orientador: Prof. Dr. Alexandre Dias Porto Chiavegatto Filho.

Versão Revisada
São Paulo
2023

GABRIEL FERREIRA DOS SANTOS SILVA

Inteligência Artificial para a vigilância de doenças crônicas não-transmissíveis

Tese apresentada ao Programa de Pós-Graduação em Saúde Pública da Faculdade de Saúde Pública da Universidade de São Paulo para obtenção do título de Doutor em Ciências.

Área de concentração: Saúde Pública.

Orientador: Prof. Dr. Alexandre Dias Porto Chiavegatto Filho.

São Paulo
2023

FOLHA DE AVALIAÇÃO

Silva GFS. **Inteligência Artificial para a vigilância de doenças crônicas não-transmissíveis**. Tese (Doutorado em Saúde Pública). Faculdade de Saúde Pública da Universidade de São Paulo, São Paulo – SP, 2023.

Aprovado em:

Banca Examinadora:

Dra. Alessandra Carvalho Goulart
Instituição: Hospital Universitário, Universidade de São Paulo
Julgamento: _____

Dr. André Filipe de Moraes Batista
Instituição: Insper Instituto de Ensino e Pesquisa
Julgamento: _____

Dr. Francisco Chiaravalloti Neto
Instituição: Faculdade de Saúde Pública, Universidade de São Paulo
Julgamento: _____

Catalogação da Publicação

Ficha elaborada pelo Sistema de Geração Automática a partir de dados fornecidos pelo(a) autor(a)
Bibliotecária da FSP/USP: Maria do Carmo Alvarez - CRB-8/4359

Silva, Gabriel Ferreira dos Santos
Inteligência Artificial para a vigilância de doenças crônicas não-transmissíveis / Gabriel Ferreira dos Santos Silva; orientador Dr. Alexandre Dias Porto Chiavegatto Filho . -- São Paulo, 2023.
100 p.

Tese (Doutorado) -- Faculdade de Saúde Pública da Universidade de São Paulo, 2023.

1. Machine Learning. 2. Doenças crônicas não transmissíveis. 3. Gestão de Saúde. 4. Saúde Pública. I. , Dr. Alexandre Dias Porto Chiavegatto Filho, orient. II. Título.

Dedico esta tese à minha família e à saúde pública do Brasil.

Agradecimentos

A Deus, por iluminar minhas decisões e abrir caminhos para a realização deste doutorado.

Aos meu pais, Adriana e José Mauro, à minha irmã, Manuela, aos meus avós, José e Maria, a Luciano Fernandes e aos meus tios, Paula e André, pelo cuidado, carinho e por toda a torcida durante minha trajetória.

À minha avó, Lucia Rosa, que infelizmente nos deixou em razão da pior crise sanitária dos últimos anos, a pandemia causada pela COVID-19.

À Janaina, minha namorada, pelo companheirismo, cuidado e amor que se estende desde a graduação e à minha sogra, Sueli, pelas orações e por toda a preocupação.

Ao professor Dr. Alexandre Chiavegatto Filho, meu orientador, por ter me guiado de maneira ímpar, contribuindo para o meu desenvolvimento como pessoa, profissional, cidadão e cientista.

À toda a equipe de pesquisadores e amigos do Laboratório de Big Data e Análise Preditiva em Saúde (LABDAPS-FSP) pelas conversas, discussões científicas, almoços, jantas, pizzas e parcerias de trabalho.

Aos professores e funcionários da Faculdade de Saúde Pública da USP pelo solícito suporte ao longo do curso e por todas as contribuições em meu processo de formação.

À equipe da Divisão de Doenças e Agravos Não-Transmissíveis da Secretaria de Estado da Saúde de São Paulo, representada pela Dra. Luciane Duarte, Dra. Mirian Shirassu e Dr. Marco Antonio, pela parceria de trabalho e pela confiança.

A todos aqueles que contribuíram, direta ou indiretamente, para a concretização desta etapa.

E ao Fundo Especial de Saúde para Imunização em Massa e Controle de Doenças (FESIMA), pelo investimento financeiro e suporte institucional ao projeto que deu origem a essa Tese de Doutorado.

Silva GFS. Inteligência Artificial para a vigilância de doenças crônicas não-transmissíveis. 2023 [tese]. São Paulo: Faculdade de Saúde Pública da Universidade de São Paulo; 2023.

Resumo

As doenças crônicas não transmissíveis (DCNT) representam um desafio significativo para a saúde global, exercendo impacto substancial nos sistemas de saúde em todo o mundo e demandando ações de vigilância e gestão. Nos últimos anos, a utilização de algoritmos de Machine Learning (ML) tem se mostrado uma abordagem promissora para aprimorar ao cuidado e a gestão de saúde. Nesse sentido, esta tese buscou desenvolver algoritmos de ML que contribuam para a vigilância, prevenção e tratamento de DCNT, com o objetivo de colaborar com a saúde pública através de dados e inteligência artificial (IA). Para isso, foram desenvolvidos, em parceria com a Secretaria de Estado da Saúde de São Paulo, quatro manuscritos com distintas aplicações, que compõem a coletânea de artigos desta tese. No primeiro artigo, foi desenvolvida uma revisão sistemática da literatura para explorar o uso de algoritmos de ML na predição da hipertensão arterial. Vinte e um artigos publicados entre janeiro de 2018 e maio de 2021 foram analisados, demonstrando o potencial dos algoritmos de ML para predizer a hipertensão e aprimorar as decisões clínicas preventivas, ainda que alguns dos trabalhos avaliados tenham apresentado problemas de seleção de variáveis e adoção de boas práticas preditivas. O segundo artigo concentrou-se na predição do risco de mortalidade em pacientes com neoplasias malignas no estado de São Paulo. Utilizando dados longitudinais, algoritmos de ML foram testados, alcançando altos valores de Área sob a curva ROC (AUC-ROC) para diferentes tipos de câncer (acima de 0,90). Os resultados apontaram para o potencial para predizer o risco de óbito em pacientes com câncer no estado de São Paulo. O terceiro artigo explorou o uso de algoritmos de ML não supervisionados para a regionalização dos municípios do estado de São Paulo com base nos perfis de morbimortalidade por DCNT. Por meio do agrupamento dos 645 municípios, o estudo identificou áreas contíguas com morbidades e mortalidades semelhantes. Esta abordagem demonstrou o potencial da utilização de ML no fornecimento de informações para o planejamento e a gestão dos sistemas de saúde. Por fim, no quarto artigo buscou-se desenvolver algoritmos de ML para a avaliação da performance da gestão de saúde crônica nos municípios do estado de São Paulo. Para isso, foram calculados os valores esperados de mortalidade prematura ajustada pela idade para cada um dos municípios no período de 2010 a 2019, a partir de um algoritmo de ML. Esses valores esperados, quando comparados com o observado nesses municípios, apontaram para a presença de casos de overachievers ou underachievers, que podem direcionar políticas de saúde e a atenção a nível estadual. As pesquisas apresentadas nesses artigos têm o potencial de contribuir para o avanço das aplicações de ML no campo da saúde pública, abrindo caminhos para estratégias mais eficazes no enfrentamento das DCNT e na promoção de saúde da população.

Palavras-chave: Doenças Crônicas Não Transmissíveis; Machine Learning; Mortalidade; Gestão de Saúde.

Silva, GFS. [Artificial Intelligence for non-communicable chronic diseases surveillance] [thesis]. São Paulo: Faculdade de Saúde Pública da Universidade de São Paulo; 2023. Portuguese.

Abstract

Chronic non-communicable diseases (NCD) pose a significant challenge for global health, exerting a substantial impact on health systems worldwide, requiring surveillance and management actions. In recent years, the use of machine learning (ML) algorithms has shown promise to improve health care and management. In this sense, this thesis sought to develop ML algorithms that contribute to the surveillance, prevention, and treatment of NCD, with the aim of collaborating with public health through data and artificial intelligence (AI). To this end, four manuscripts with different applications were developed, in partnership with the São Paulo State Department of Health, which make up the collection of articles for this thesis. In the first article, a systematic literature review was developed to explore the use of ML algorithms in the prediction of arterial hypertension. Twenty-one articles published between January 2018 and May 2021 were analyzed, demonstrating the potential of ML algorithms to predict hypertension and improve preventive clinical decisions, although some of the studies evaluated presented problems of variable selection and adoption of good predictive practices. The second article was focused on predicting the risk of mortality in patients with malignant neoplasms in the state of São Paulo, Brazil. Using longitudinal data, several ML algorithms were tested, achieving high values of Area Under the ROC Curve (AUC-ROC) for different types of cancer (above 0.90). The results highlighted the potential to predict the risk of death in cancer patients in the state of São Paulo. The third article explored the use of unsupervised ML algorithms for the regionalization of municipalities in the state of São Paulo based on morbidity and mortality profiles due to NCD. By grouping the 645 municipalities, the study identified contiguous areas with similar morbidities and mortality. This approach demonstrated the potential of using ML in providing information for the planning and management of health systems. Finally, the fourth article sought to develop ML algorithms to support the evaluation of the performance of chronic health management in the municipalities of the state of São Paulo. To this end, we calculated expected values of age-adjusted premature mortality for each of the municipalities in the period from 2010 to 2019, from a ML algorithm. These expected values, when compared with those observed in these municipalities, indicate cases of overachievers or underachievers, which can guide the direction of health policies and care at the state level. The research presented in these articles contributes to the advancement of ML applications in the field of public health, opening paths for more effective strategies in coping with NCD and in promoting the health of the population.

Keywords: Chronic Non-Communicable Diseases; Machine Learning; Mortality; Health Management.

Sumário

1. Introdução 10
2. Apresentação 14
3. Material e Métodos 16
4. Resultados 16
Artigo 1: Machine learning for hypertension prediction: A systematic review 17
Artigo 2: Machine learning for longitudinal mortality risk prediction in patients with malignant neoplasm in São Paulo, Brazil 38
Artigo 3: Unsupervised machine learning for the regionalization of healthcare management in noncommunicable diseases 58
Artigo 4: Improving Health Management of Chronic Non-Communicable Diseases: Machine Learning Algorithms for Surveillance and Performance Evaluation in Healthcare Systems 75
5. Considerações Finais e Conclusão 95
Referências 99

Introdução

As Doenças Crônicas Não Transmissíveis (DCNT) são condições de saúde de longa duração e progressão gradual, geralmente resultantes da interação de múltiplos fatores, como predisposição genética, estilo de vida e hábitos alimentares [1]. Essas doenças incluem problemas como diabetes, câncer, doenças cardiovasculares e doenças respiratórias crônicas. De acordo com dados da Organização Mundial da Saúde (OMS), as DCNT representam uma parcela substancial da carga global de doenças, sendo responsáveis por aproximadamente 74% de todas as mortes no mundo, anualmente [2]. Na região das Américas, as DCNT foram responsáveis por 81% do total de óbitos em 2019 [3]. No Brasil, segundo dados do Sistema de Informações sobre Mortalidade – Ministério da Saúde (SIM/MS), as doenças cardiovasculares, os cânceres, a diabetes e as doenças respiratórias crônicas foram a causa básica de 54,3% do total de óbitos em 2019 [4].

O enfrentamento das DCNT requer uma abordagem abrangente, que envolva não apenas a detecção precoce e o tratamento adequado, mas também medidas preventivas integradas aos sistemas de saúde, desde o cuidado ao paciente até a vigilância e gestão. O Plano de Doenças e Agravos Não Transmissíveis 2022-2030, elaborado pelo Ministério da Saúde, menciona quatro eixos principais no combate às DCNT: promoção à saúde, atenção integral à saúde, vigilância em saúde e prevenção de doenças e agravos à saúde [5]. Em linhas gerais, destaca-se a educação pública, o desenho e a mensuração de políticas de saúde preventivas, a promoção de alimentação balanceada, o incentivo à prática de atividades físicas e ações para redução do tabagismo e alcoolismo como algumas das estratégias cruciais para enfrentar este desafio, especialmente na faixa etária prematura, que se estende dos 30 aos 69 anos [5].

Nessa linha de combate às DCNT, a vigilância surge como um conjunto coordenado de práticas com o propósito de compreender, antecipar, evitar e enfrentar questões relacionadas à saúde da população em um determinado território, abrangendo fatores de risco atuais e futuros, incidentes, incapacidades, enfermidades e ameaças à saúde [5].

Diante desse cenário, a Inteligência Artificial (IA), mais especificamente a área de Machine Learning (ML), apresenta-se como um conjunto de técnicas e ferramentas promissoras no apoio à vigilância de DCNT, oferecendo o potencial de aprimorar a gestão e prevenção dessas doenças. ML é a maior área da IA e se baseia no desenvolvimento e na aplicação de algoritmos que aprendem a partir de dados, identificando padrões gerais e fornecendo insumos para a tomada de decisão [6]. A capacidade de ML de processar grandes volumes de dados, identificar padrões ocultos e gerar predições tem se mostrado

relevante em diversas áreas do conhecimento. Na saúde, em função da sua complexidade, sensibilidade e criticidade dos desfechos, incluindo os relacionados às DCNT, os avanços científicos têm permitido o desenvolvimento da área e a compreensão gradual dos limites e desafios para a implementação prática desses algoritmos, como questões éticas e operacionais [7,8].

Dentre as abordagens de ML, as duas estratégias mais utilizadas são o aprendizado supervisionado e o não supervisionado. O aprendizado supervisionado envolve o treinamento de algoritmos com um conjunto de dados rotulados, ou seja, dados que contêm informações sobre as características dos pacientes e as respostas esperadas, como diagnósticos ou classificações de risco [9,10]. Nesse tipo de aprendizado, o objetivo é que o algoritmo seja capaz de mapear os inputs (preditores) e atribuir um output correto (predição), permitindo a antecipação de novos casos e a tomada de decisão clínica baseada em evidências.

Por outro lado, o aprendizado não supervisionado não requer dados rotulados para o treinamento do algoritmo. Nessa abordagem, os dados são analisados sem categorias ou rótulos prévio, buscando-se identificar padrões e estruturas intrínsecas aos dados, geralmente não visíveis ao olho humano [11,12].

Dada a complexidade e a natureza multifatorial das DCNT [13], a utilização de ML pode proporcionar avanços na identificação precoce de riscos, no aprimoramento do manejo clínico, no suporte à formulação de políticas de saúde pública e na avaliação da performance do sistema de saúde. Nesse sentido, esta tese de doutorado buscou utilizar aplicações de ML, tanto no aprendizado supervisionado, quanto no não supervisionado, para contribuir com o enfrentamento das DCNT e auxiliar na promoção da saúde e no combate a essas enfermidades crônicas de alto impacto. Dessa forma, através de quatro manuscritos, buscou-se atingir os seguintes objetivos:

Objetivo principal:

- Desenvolver modelos de ML para suporte à tomada de decisão a nível de gestão das políticas de saúde, fornecendo insumos para os órgãos governamentais de vigilância em DCNT.

Objetivos secundários:

- Identificar o estado da arte, desafios e oportunidades nas aplicações de ML para predição de hipertensão arterial sistêmica, um importante desfecho crônico

cardiovascular.

- Desenvolver, de forma científica, crítica e criteriosa, a área de ML em saúde, com um foco específico no enfrentamento às DCNT. Ao fortalecer o desenvolvimento teórico e crítico, a pesquisa busca contribuir diretamente com informações e modelos para a vigilância em saúde, melhorando a interpretação e implementação de ferramentas baseadas em ML na gestão de DCNT.
- Desenvolver modelos de ML capazes de orientar a tomada de decisão clínica, especialmente em desfechos relacionados às DCNT. Essa iniciativa visa a criação de ferramentas práticas para profissionais de saúde e órgãos de vigilância, proporcionando insights preditivos em tempo real.

A prevalência/incidência e o impacto das DCNT na saúde pública são desafios complexos e multifatoriais, demandando abordagens inovadoras para aprimorar a identificação precoce de riscos, otimizar o manejo clínico, formular políticas de saúde pública eficazes e avaliar a performance do sistema de saúde. Nesse contexto, a aplicação de técnicas de ML emerge como uma ferramenta promissora, oferecendo oportunidades significativas para avançar no enfrentamento dessas enfermidades crônicas de alto impacto.

A presente tese de doutorado tem como premissa fundamental a utilização de aplicações de ML, tanto no aprendizado supervisionado quanto no não supervisionado, para contribuir de maneira substancial no combate às DCNT. O objetivo principal é o desenvolvimento de modelos de ML voltados para a gestão das políticas de saúde, fornecendo insumos cruciais para os órgãos governamentais de vigilância em DCNT. Esse enfoque estratégico visa aprimorar a eficácia das intervenções, promovendo uma abordagem mais assertiva e personalizada na prevenção, no tratamento e na gestão das DCNT.

Os objetivos secundários desta pesquisa buscam complementar e aprofundar a contribuição para a área. A realização de uma revisão sistemática da literatura, com enfoque na hipertensão, visa identificar o estado da arte das aplicações de ML na predição desta importante doença cardiovascular, proporcionando uma visão abrangente das lacunas existentes.

A criação de modelos de ML destinados a orientar a tomada de decisão clínica em desfechos relacionados às DCNT representa um avanço importante na interface entre a tecnologia e a prática clínica. Ao integrar dados complexos e variáveis, esses algoritmos

têm o potencial de fornecer insights e informações de risco, apoiando os profissionais de saúde na tomada de decisões mais informadas e personalizadas, com o potencial de promover a melhora nos desfechos de saúde e, consequentemente, na gestão e na vigilância de DCNT.

Apresentação

Diante dos objetivos apresentados, esta tese foi estruturada em quatro capítulos, buscando-se desenvolver distintas ferramentas de ML e demonstrar a capacidade de implementação desses algoritmos na prática clínica e na vigilância em saúde. Além dos objetivos gerais da tese, cada capítulo possui seus objetivos específicos, apresentados individualmente na seção de resultados.

O primeiro artigo foi uma revisão sistemática da literatura publicada no periódico *Current Hypertension Reports* [14], analisando estudos que desenvolveram algoritmos de ML para a predição da hipertensão arterial primária. Por meio de uma metodologia baseada no Transparent Reporting of Systematic Reviews and Meta-Analyses (PRISMA) [15] e com o auxílio de ferramentas de IA para triagem de artigos científicos [16], foram identificados 21 trabalhos publicados entre janeiro de 2018 e maio de 2021, que desenvolveram métodos para a predição de hipertensão por meio de diversos algoritmos, como o Support Vector Machine (SVM), o Extreme Gradient Boosting (XGBoost) e Random Forest.

Essa análise inicial proporcionou uma visão do potencial dessas técnicas de ML para aprimorar a prevenção e o manejo clínico da hipertensão. No entanto, também foi possível identificar a necessidade de uma abordagem crítica, criteriosa e científica, considerando fatores técnicos como a definição clara dos desfechos, a interpretabilidade dos modelos, a padronização das métricas de desempenho apresentadas e os possíveis vazamentos de informação (data leakage).

No segundo artigo, foram desenvolvidas aplicações de ML para a predição do risco de mortalidade em pacientes com neoplasias malignas no estado de São Paulo, Brasil. Para isso, foram utilizados dados do Registro Hospitalar de Câncer da Fundação Oncocentro de São Paulo (RHC-FOSP), que, dada sua característica longitudinal de acompanhamento e coleta, permitiu o desenvolvimento de algoritmos com boa performance preditiva.

Os algoritmos utilizados alcançaram valores elevados de Área sob a curva ROC (AUC-ROC) na predição do risco de morte para diferentes tipos de câncer. Esses achados proporcionam uma base sólida para demonstrar o potencial de incorporação destas ferramentas no suporte às decisões clínicas em pacientes oncológicos do estado de São Paulo. Esse artigo foi publicado na revista *Artificial Intelligence in the Life Sciences* [17].

O terceiro artigo concentra-se na aplicação de algoritmos de ML não supervisionados para regionalizar os municípios do estado de São Paulo, Brasil, segundo o perfil de morbimortalidade por DCNT. Por meio de técnicas de clustering, como o SKATER e k-

means, foram identificadas áreas com perfis epidemiológicos e locais semelhantes, fornecendo uma ferramenta para o planejamento e gerenciamento do sistema de saúde do Estado de São Paulo. Esse artigo foi submetido para publicação e encontra-se em processo de avaliação.

No quarto artigo, foram desenvolvidos algoritmos de ML para auxiliar na avaliação do desempenho da gestão de doenças crônicas nos municípios do estado de São Paulo. Para isso, algoritmos de ML estimaram as taxas de mortalidade prematura ajustadas pela idade em cada município durante o período de 2010 a 2019. Essas estimativas, quando contrastadas com os números reais observados em cada município, identificaram situações de alto valor ou baixo valor em relação ao esperado para determinada cidade, dadas suas características gerais.

Após a apresentação dos quatro artigos, a seção de Considerações Finais e Conclusão apresenta uma retomada dos trabalhos desenvolvidos, destacando algumas das limitações identificadas e ampliando os horizontes para trabalhos futuros.

Esta tese é fruto de uma colaboração entre a Secretaria de Estado da Saúde de São Paulo (SES/SP), mais especificamente da Divisão de Doenças Crônicas não Transmissíveis do Centro de Vigilância Epidemiológica "Prof. Alexandre Vranjac", e o Laboratório de Big Data e Análise Preditiva em Saúde da Faculdade de Saúde Pública da USP (LABDAPS/FSP). A parceria entre o meio acadêmico e o órgão governamental proporcionou o desenvolvimento de aplicações de ML na gestão de saúde, com o objetivo de melhorar a saúde pública e os cuidados em DCNT.

Este trabalho foi financiado pelo Fundo Especial de Saúde para Imunização em Massa e Controle de Doenças (FESIMA), o que possibilitou o desenvolvimento das análises. A parceria com o FESIMA e a SES/SP foi fundamental para viabilizar a pesquisa e potencializar o impacto das descobertas no contexto da saúde pública, permitindo o desenvolvimento de soluções práticas e aplicáveis para o enfrentamento das DCNT no Estado de São Paulo.

O projeto foi aprovado pelo Comitê de Ética em Pesquisa da Faculdade de Saúde Pública da USP, sob o CAAE: 65375722.9.0000.5421. Os materiais suplementares respectivos aos artigos estão disponíveis no GitHub[1].

[1] https://github.com/gabriel1710/tese_material_suplementar/tree/master

Material e Métodos

A presente tese foi estruturada sob a forma de coletânea composta por quatro artigos científicos, uma abordagem em conformidade com as diretrizes estabelecidas no Regulamento do Programa Pós-graduação em Saúde Pública (2023), especificamente no item XI.2 Formato das Teses de Doutorado, conforme instituído pela Resolução CoPGr nº 8376, datada de 7 de março de 2023. Essa estrutura permite uma abordagem mais específica e aprofundada dos materiais e métodos utilizados em cada artigo, oferecendo uma visão detalhada das abordagens adotadas em diferentes contextos. Cada artigo apresentará de maneira individualizada os procedimentos, técnicas e análises empregados, proporcionando uma compreensão clara e aprofundada das contribuições específicas de cada componente desta coletânea científica.

Resultados

Os resultados provenientes desta pesquisa referem-se aos quatro artigos científicos apresentados nas seções a seguir.

Artigo 1: Machine learning for hypertension prediction: A systematic review

Current Hypertension Reports (2022) 24:523–533
https://doi.org/10.1007/s11906-022-01212-6

GUIDELINES/CLINICAL TRIALS/META-ANALYSIS (WJ KOSTIS, SECTION EDITOR)

Machine Learning for Hypertension Prediction: a Systematic Review

Gabriel F. S. Silva[1] · Thales P. Fagundes[2] · Bruno C. Teixeira[2] · Alexandre D. P. Chiavegatto Filho[1]

Accepted: 8 June 2022 / Published online: 22 June 2022

Abstract

Purpose of Review To provide an overview of the literature regarding the use of machine learning algorithms to predict hypertension. A systematic review was performed to select recent articles on the subject.

Recent Findings The screening of the articles was conducted using a machine learning algorithm (ASReview). A total of 21 articles published between January 2018 and May 2021 were identified and compared according to variable selection, train-test split, data balancing, outcome definition, final algorithm, and performance metrics. Overall, the articles achieved an area under the ROC curve (AUROC) between 0.766 and 1.00. The algorithms most frequently identified as having the best performance were support vector machines (SVM), extreme gradient boosting (XGBoost), and random forest.

Summary Machine learning algorithms are a promising tool to improve preventive clinical decisions and targeted public health policies for hypertension. However, technical factors such as outcome definition, availability of the final code, predictive performance, explainability, and data leakage need to be consistently and critically evaluated.

Keywords Hypertension · Machine learning · Systematic review · Evaluation metrics · Model construction

ABSTRACT

Purpose of Review To provide an overview of the literature regarding the use of machine learning algorithms to predict hypertension. A systematic review was performed to select recent articles on the subject.

Findings The screening of the articles was conducted using a machine learning algorithm (ASReview). A total of 21 articles published between January 2018 and May 2021 were identified and compared according to variable selection, train-test split, data balancing, outcome definition, final algorithm, and performance metrics. Overall, the articles achieved an Area Under the ROC Curve (AUROC) between 0.766 and 1.00. The algorithms most frequently identified as having the best performance were Support Vector Machines (SVM), Extreme Gradient Boosting (XGBoost), and random forest.

Summary Machine learning algorithms are a promising tool to improve preventive clinical decisions and targeted public health policies for hypertension. However, technical factors such as outcome definition, availability of the final code, predictive performance, explainability and data leakage need to be consistently and critically evaluated.

Keywords: hypertension; machine learning; systematic review; evaluation metrics; model construction.

INTRODUCTION

Systemic Arterial Hypertension (SAH) is a chronic disease that affects more than one billion patients worldwide and is present in one in four men and one in five women [1]. Essential hypertension, also called primary hypertension, is associated with increased blood pressure caused by several factors such as genetic mutations and polymorphisms, high salt and alcohol intake, aging, and sedentary lifestyle, and increases the risk of developing kidney, heart, and brain diseases [2–4]. On the other hand, secondary hypertension is associated with elevated blood pressure from known and generally reversible causes, representing about 5% to 10% of cases [5, 6].

Notably, from the 2000s onwards science has been increasingly influenced by the era of big data [7], which has allowed for the emergence of machine learning applications. In general, machine learning is a set of data-driven tools used to support decision-making. It has had recent applications in many scientific areas, especially in healthcare. It has shown potential in predicting health outcomes, from diseases and injuries to even deaths [8]. Supervised learning algorithms are the most frequent of machine learning applications, where algorithms learn to predict a specific outcome, such as the incidence of diseases or future patient prognosis [8].

Considering the recent growth in machine learning applications for healthcare, it is important to identify current trends in the field in order to fill relevant gaps in the literature and to make concrete progress on the subject. This study aims to (1) critically assess the recent literature regarding the application of machine learning in predicting the incidence of primary hypertension; (2) highlight the most critical findings and describe the landscape of the field; (3) identify the most popular algorithms for predicting hypertension; (4) present a synthesis of the published literature; and (5) identify the opportunities and future paths for applying machine learning algorithms to predict hypertension.

MATERIAL AND METHODS

We conducted a systematic review following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) recommendations. Two authors (TPF e GFSS) independently searched MEDLINE (through PubMed), Embase, ClinicalTrials.gov, and Web of Science for studies published from January 2018 to May 2021. We included search strings including keywords such as 'Machine Learning', 'Artificial Intelligence', 'Deep Learning' + 'Essential Hypertension', 'Hypertension', 'Arterial Hypertension'. The complete list of strings used can be found in the Table 1.

Table 1. Search strings used to survey the literature. (To be continued)

Database	Machine Learning rules	Hypertension rules
clinicaltrials.gov	Artificial Intelligence	Arterial, hypertension
Embase	('artificial intelligence'/exp OR 'artificial intelligence' OR 'machine learning'/exp OR 'machine learning' OR 'deep learning'/exp OR 'deep learning' OR 'supervised machine learning'/exp OR 'supervised machine learning' OR 'unsupervised machine learning'/exp OR 'unsupervised machine learning' OR 'natural language processing'/exp OR 'natural language processing' OR 'neural networks, computer'/exp OR 'neural networks, computer' OR 'computer reasoning'/exp OR 'computer reasoning' OR 'knowledge acquisition (computer)' OR 'machine intelligence'/exp OR 'machine intelligence' OR 'ai (artificial intelligence)' OR 'computer vision systems')	('essential hypertension'/exp OR 'essential hypertension' OR 'renovascular hypertension'/exp OR 'renovascular hypertension')

Table 1. Search strings used to survey the literature. (Conclusion)

Database	Machine Learning rules	Hypertension rules
PubMed	("Artificial Intelligence"[MeSH] OR "machine learning"[MeSH] OR "deep learning"[MeSH] OR "Supervised Machine Learning"[MeSH] OR "Unsupervised Machine Learning"[MeSH] OR "Natural Language Processing"[MeSH] OR "Neural Networks, Computer"[MeSH] OR "Computer Reasoning" OR "Machine Intelligence" OR "AI (Artificial Intelligence)" OR "Computer Vision Systems")	(("Essential Hypertension" OR "systemic arterial hypertension" OR "arterial hypertension" OR "hypertension") NOT ("portal hypertension" OR "pulmonary hypertension" OR "Pulmonary arterial hypertension"))
Web of Science	("Artificial Intelligence" OR "machine learning" OR "deep learning" OR "Supervised Machine Learning" OR "Unsupervised Machine Learning" OR "Natural Language Processing" OR "Neural Networks, Computer" OR "Computer Reasoning" OR "Machine Intelligence" OR "AI (Artificial Intelligence)" OR "Computer Vision Systems")	("Essential Hypertension" OR "systemic arterial hypertension" OR "arterial hypertension" OR "hypertension" NOT "portal hypertension" OR "pulmonary hypertension" OR "Pulmonary arterial hypertension")

The entries were exported to a spreadsheet, and duplicates were removed. The reference lists from every individual study were also manually searched. Each study was then independently reviewed by three of the authors to confirm if it was within the inclusion criteria. All the selected articles were evaluated to identify the risk of biases and the possibility of data leakage.

Inclusion Criteria

The articles were included in this review according to the following criteria: 1) articles that effectively created and tested machine learning models, presenting the performance criteria for the algorithms; 2) articles focused on primary hypertension; 3) articles that assessed the stratification of the risk of developing primary hypertension. 4) articles that used either separation of the data set in training and testing or cross-validation.

Exclusion Criteria

Articles that presented the following characteristics were excluded from the review: 1) literature review articles; 2) articles that did not specify the algorithms used; 3) articles

that evaluated secondary outcomes of hypertension, such as chronic renal failure, stroke, acute myocardial infarction, among others; 3) articles without information regarding the outcome; 4) articles that used only genetic data for classification; 5) articles that assessed outcomes of secondary hypertension or other specific types of hypertension, such as intracranial and pulmonary; 6) articles that only included pregnant women and children.

Outcome

A positive diagnosis for systemic arterial hypertension was the only outcome analyzed. The effect measure used to synthesize the results was the Area Under the ROC Curve (AUROC), which is currently the most frequent metric used for classification algorithms.

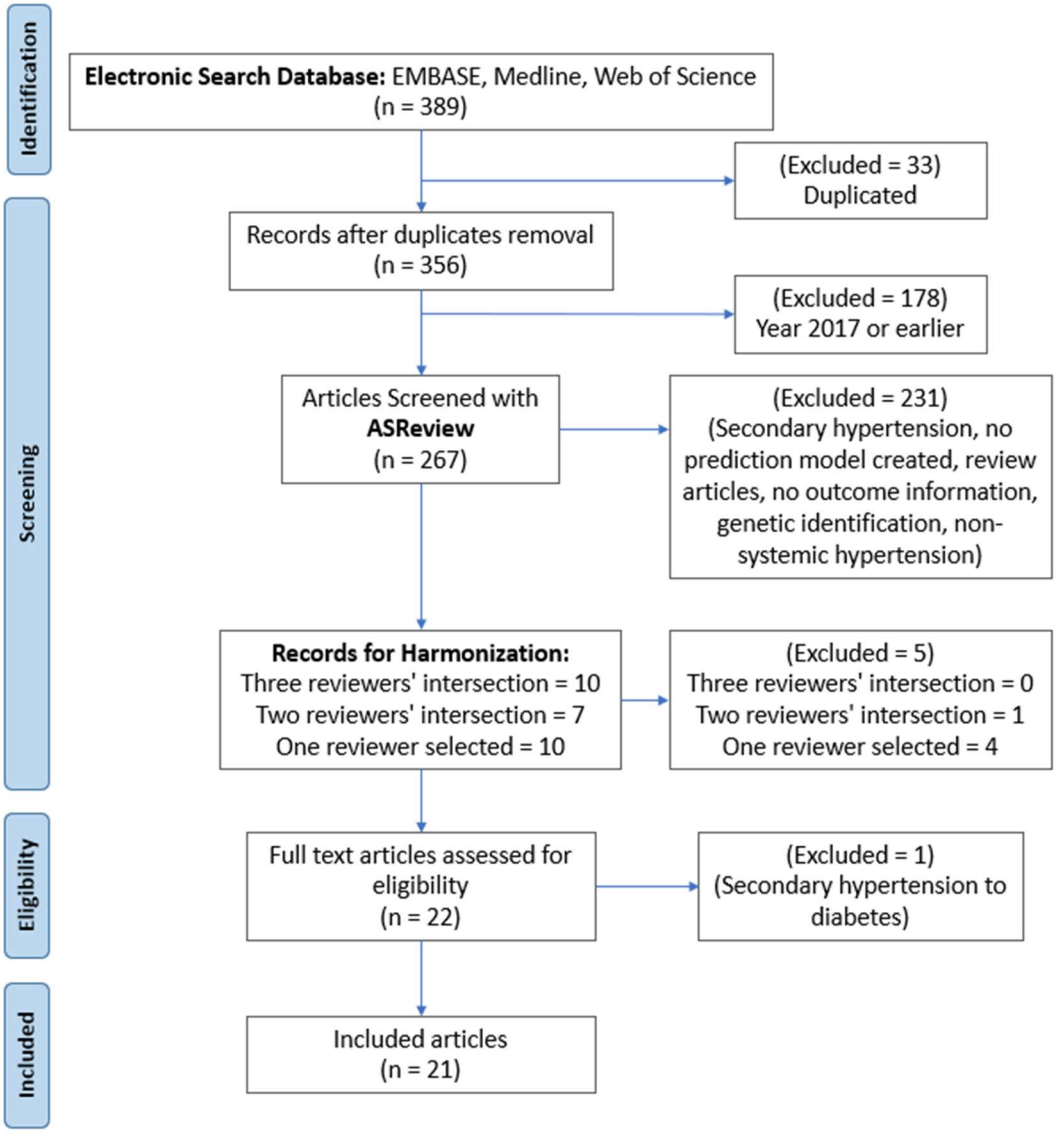
Screening and Selection Strategy

In order to perform the screening of the literature, we used a machine learning algorithm (ASReview) [9] to rank articles based on their textual proximity with previously selected articles, reducing the time and effort expenditure in the early screening phase. This process was carried out independently by three authors. The selected articles were then compared to assess a) which were commonly selected by all, b) which were at the intersection between two authors, c) and which were selected by only one author. The articles selected by the three authors were automatically included in the review, and the remaining articles (intersection between two authors or individual selection) had their inclusions re-evaluated by consensus. The selection of articles was therefore performed by following these steps: (1) screening with ASReview, (2) contrast of individual results, (3) reassessment of articles that were not commonly selected by the three authors, and (4) final selection.

RESULTS

Using an article screening tool based on machine learning considerably reduced the manual review process by efficiently selecting and excluding articles after the first literature search. The final selection was obtained according to the workflow presented in Figure 1.

Figure 1. PRISMA procedure workflow.



Overview of individual studies

The datasets of the studies presented great diversity in terms of sample size, nationality, and availability of variables. Of the 21 articles selected, seven (33.3%) used Chinese data, five (23.8%) American, two (9.5%) Indian, and two (9.5%) South Korean. The other datasets came from Brazil, Italy, Canada, Singapore, Japan, and Qatar (4.8% for each

country). The features available in these datasets are diverse, ranging from clinical, geographic, socioeconomic, nutritional, and routine data to genetic data.

Relationship between sample size and algorithm performance

In terms of sample size, the articles were classified into 6 categories: 139-300 observations (3 articles, 14.3%), 301-1,000 observations (5 articles, 23.8%), 1,001-5,000 observations (5 articles, 23.8%), 5,001-10,000 observations (3 articles, 14.3%), 10,001 – 100,000 observations (4 articles, 19.0%) and 100,000+ observations (1 article – 4.8%). There was no significant correlation between the sample size and the Area Under ROC Curve (AUROC), with a Pearson's product-moment correlation of 0.258 (p-value = 0.3531). Regarding the programming language, R was used in six studies, Python in five, and Waikato Environment for Knowledge Analysis (WEKA) was used in three. Other tools like Matlab and TensorFlow were seen in at least one article. Three articles did not specify the language used to build the models. Only one study shared the code used for analysis, and another mentioned it was available upon request.

Model Construction

The construction of the models was evaluated according to five elements: feature selection, train-test split, data balancing, outcome definition, final algorithm, and performance metrics. Feature selection is an important process during the modeling process to remove irrelevant or unimportant information in the construction of the model, helping to decrease the risk of overfitting [10]. Of the 21 articles, five of them did not perform or did not inform the variable selection strategy. Six studies performed the selection based on the risk factors already reported in previous studies, and one conducted a significance analysis of the variables. Additionally, three articles used the random forest algorithm for variable selection, three the information gain-based feature selection, and one employed logistic regression. The other articles individually used the chi-squared test, Empirical Mode Decomposition (EMR), XGBoost, genetic algorithm-based feature selection, Correlation-based Feature Selection (CFS), Wrapper-based Feature Selection (WFS), and Minimum-Redundancy-Maximum-Relevance (mRMR) for feature selection. Among the articles that applied the variable selection strategy, eleven used one strategy, three used two, and one tested five different strategies. Feature selection was not performed in one article as it analyzed only image data.

Regarding the train-test split, seven articles used k-fold cross-validation and ten studies performed a full split strategy between training and testing, with training percentages ranging between 57.65% and 90%, and testing between 10% and 42.35%. Two of them did not specify the percentage allocated between training and testing datasets. One study used the training, testing, and validation approach, with a proportion of 80%, 10%, and 10%, respectively. Two articles performed simulations to assess which strategy offered the best performance, testing at least two different splits for training and testing. One article applied four different splits strategies: 60% training, 40% testing; 70% training, 30% testing; 80% training, 20% testing; 90% training, 10% testing [11]. Two articles used both cross-validation and split training and testing, with one of them comparing the results between cross-validation and split into training and test [12], while the other, despite presenting the results of both approaches, used cross-validation for hyperparameters tunning [13].

Outcome classes

Binary outcomes, i.e., “normotensive” vs. “hypertensive”, was the most common approach for outcome selection (19 studies). Two articles evaluated multiple outcomes: one classified the outputs between “normal”, “pre-hypertension” and “hypertension” [14], while the other classified the outcome as “normal”, “Prehypertension”, “stage-1 hypertension” and “stage-2 hypertension” [15]. The definition of hypertension was mostly based on Systolic Blood Pressure (SBP) values greater than 130-140 and Diastolic Blood Pressure (DPB) greater than or 90 mm Hg. Three articles did not clearly specify the hypertensive outcome.

Most articles worked with unbalanced datasets regarding the outcome variable. Two of the 21 articles did not report the distribution of the outcome. Of the 18 that presented the balance, the positive class (hypertensive individuals) ranged from 6.68% to 88.5%.

Regarding the best-performing algorithms, five articles selected the Support Vector Machines (SVM) algorithm as the final model, four the Random Forest, two the XGBoost, two the K-NN, and two the Artificial Neural Networks (ANN). The other five articles reported as the best performing algorithm the C4.5, Convolutional Neural Networks, Logistic Regression, Cox regression, and Naive Bayes. An ensemble method was considered the best performing algorithm in one article.

The Area Under the ROC Curve (AUROC) was the more frequently evaluated performance metric on the test set or by cross-validation. Six articles, however, did not present the result for the AUROC, preferring other metrics such as accuracy, precision, recall, sensitivity, specificity, and/or F1-score. Overall, the articles found a high AUROC,

ranging from 0.766 to 1.00. Three articles found an AUROC equal or greater than 0.90, which is considered a very high predictive performance, despite the possibility of bias and data leakage.

For a more in-depth analysis, Table 2 presents a series of metrics from the identified articles, with the description of items such as sample size, features, training and test strategy, balancing, among others.

Table 2. Summary information of the selected articles. (To be continued)

Article	Database	Sample Size	Features	Training and Validation Strategy	Balancing	Features Selection Strategy	Best Algorithm	AUC	System Used	Number of Citations (Scholar, until 15 Aug 2021)
[16]	Maine Health Information Exchange Network (USA)	823,627 in a retrospective cohort. 680,810 in a prospective cohort.	80 features, between age, gender, diseases, medications of depression, medication of anxiety, medication of schizophrenia, medical visits, social determinants, medication of lipid disorders, medication of type 2 diabetes, medication of cardiovascular diseases	One cohort for train and the other for validation	Hypertensive: 11.2% Normal: 88.8%	XGBoost	XGBoost	0.917	Non-Specified	62
[17]	Data from a private university in Vitória da Conquista, Bahia, Brazil	155	Obesity, waist-hip ratio, hip circumference, BMI, waist circumference, age, subject ID	10-fold Cross-Validation	Hypertensive: 23.9% Normal: 76.1%	Information gain-based feature selection	Random Forest	NA	WEKA	56
[12]	Henry Ford Health Systems (USA)	23,095	Age, METs, resting SBP, peak DBP, resting DBP, HX coronary artery disease, reason for test, history of diabetes, percentage HR achieved, race, history of hyperlipidemia, aspirin use, hypertension response	10-fold Cross-Validation, 80% - train and 20%-test, and 70% - train and 30%-test	Hypertensive: 35.0% Normal: 65.0%	Information gain-based feature selection	Random Forest	0.880 (test result under 80/20 split)	WEKA and R	49

Table 2. Summary information of the selected articles. (Continued)

Article	Database	Sample Size	Features	Training and Validation Strategy	Balancing	Features Selection Strategy	Best Algorithm	AUC	System Used	Number of Citations (Scholar, until 15 Aug 2021)
[18]	Singapore Epidemiology of Eye Disease (SEED)	2,705	SBP, age, HBA1C, arteriolar vessel caliber, venular vessel caliber, retinopathy, anti-diabetes, BMI, gender, income $2000-3000, DBP, income >3000, past smoker, alcohol, education, current smoker, Indian nationality, income $1000-2000, hyperlipidemia, Malay's ethnicity, anti-cholesterol, education (university level)	70% - train 30% - test	Hypertensive: 34.6% Normal: 65.4%	Absolute z-statistic from Logistic Regression	Support Vector Machine	0.780	R	38
[19]	Epidemiological investigation questionnaire from Beijing Chinese Han population	1,200	Environmental factors and Genetic Factors	10-fold Cross-Validation	Hypertensive: 46.6% Normal: 53.4%	Literature	Support Vector Machine	0.886	R	17
[20]	Japan Health Promotion Foundation	18,258	SBP at Year (−1), CAVI measurement Clinic, SBP at Year (−1), DBP at Year (−1), CAVI measurement SBP at Year (−2), CAVI measurement Clinic SBP at Year (−2), Clinic DBP at Year (−2), DBP at Year (−2), CAVI measurement Clinic DBP at Year (−1), BMI at Year (−1), Age at Year (−2), BMI at Year (−2), Age at Year (−1), CAVI at Year (−2), Clinic SBP by SBP at CAVI measurement at Year (−1), Waist at Year (−1), Triglycerides at Year (−2), Clinic DBP by DBP at CAVI measurement at Year (−1), CAVI at Year (−1), ALP at Year (−1), Fasting glucose at Year (−2)	75% derivation 25% validation	Hypertensive: 14.6% Normal: 85.4%	Uninformed	Ensemble	0.881	R	15

Table 2. Summary information of the selected articles. (Continued)

Article	Database	Sample Size	Features	Training and Validation Strategy	Balancing	Features Selection Strategy	Best Algorithm	AUC	System Used	Number of Citations (Scholar, until 15 Aug 2021)
[21]	Massachusetts Institute of Technology-Beth Israel Hospital (MIT-BIH, USA) and Smart Health for Assessing the Risk of Events via ECG (SHAREE, Italy)	139	ECG Signals	10-fold Cross-Validation	Hypertensive: 88.5% Normal: 11.5%	Empirical Mode Decomposition	KNN	NA	Matlab	13
[11]	GlaxoSmithKline Research genetic and genomic research (Toronto, Canada)	498	SBP, gender, age, BMI, smoking status, exercise level, alcohol consumption level, stress level, and salt intake level	Train-Test 60%-40% 70%-30% 80%-20% 90%-10%	Uninformed	Literature	Artificial Neural Network	NA	Matlab	12
[22]	6-year population-based prospective cohort study in the rural areas of Henan Province, China	8,319	Demographic characteristics and Biochemical indexes	57,65% - Train 42,35% - Test	Hypertensive: 21.6% Normal: 78,4%	Uninformed	Cox Regression	Men: 0.771 Women:0.765	R	10

Table 2. Summary information of the selected articles. (Continued)

Article	Database	Sample Size	Features	Training and Validation Strategy	Balancing	Features Selection Strategy	Best Algorithm	AUC	System Used	Number of Citations (Scholar, until 15 Aug 2021)
[14]	Korea National Health and Nutrition Examination Survey	8,212	Height, weight, waist circumference, waist-to-height circumference ratio, BMI, glucose, HBA1C, total cholesterol, HDL, triglyceride, aspartate aminotransferase, alanine aminotransferase, hemoglobin, hematocrit, BUN, CRT, WBC, RBC, FVC, FVCP, FEV1, FEV1P, FEV1FVC, FEV6, FEF25-75, PEF.	10-fold Cross Validation	Hypertensive: 38.6% Pre-hypertensive: 24.4% Normal: 37.0%	correlation-based feature selection (CFS) and wrapper-based feature selection (WFS) methods	Logistic Regression	0.845	WEKA	10
[23]	Beijing Anzhen Hospital, Capital Medical University, Beijing, China.	965	Anthropometry, personal, clinical, and genetic data	Train and test split. Percentages were not informed.	Hypertensive: 39.0% Normal: 61.0%	Literature	Support Vector Machine	SBP: 0.673 DBP: 0.817	R	9
[24]	Rural areas of Xinxiang County, Henan, in central China	625	Retinal fundus image	80% - train 10% - validation 10% - test	Hypertensive: 40.5% Normal: 59.5%	Not applicable	Convolutional Neural Network	0.766	TensorFlow	9
[25]	National Health and Nutrition Examination Survey (USA)	24,434	Race, age, smoking, BMI, diabetes, and kidney conditions.	70% - train 30% - test	Hypertensive: 30.1% Normal: 69.9% (test dataset)	Literature and p-value	Artificial Neural Network + SMOTE	0.77	Python	7

Table 2. Summary information of the selected articles. (Continued)

Article	Database	Sample Size	Features	Training and Validation Strategy	Balancing	Features Selection Strategy	Best Algorithm	AUC	System Used	Number of Citations (Scholar, until 15 Aug 2021)
[15]	Clinical data of patients admitted to the Guilin People's Hospital in Guilin, China	219	Sex, age, height (cm), weight (kg), SBP, DBP, HR, and BMI	5-fold cross-validation	Hypertensive-stage 1: 15.5% Hypertensive-stage 2: 9.2% Pre-hypertensive: 38.8% Normal: 36.5%	information gain-based feature selection and genetic algorithm-based feature selection	C4.5	1	Python	6
[26]	Electronic Medical Records from Tamil Nadu Health System Project	599	Behavioral: Smoking, Smoking frequency, Type of tobacco, alcohol consumption, alcohol frequency, smokeless tobacco, smokeless tobacco frequency, diet, non-veg frequency, oil used, physical activity, duration Medical: History of HTN, Family history of HTN, complication of HTN, history of diabetes, family history of diabetes, symptoms of diabetes, complication of diabetes, history of other disorder	70% - train 30% - test	New Diagnostics: 3.3% Known Hypertensive: 3.3% Pre-Hypertensive: 53.6% Normal: 39.8%	Uninformed	Support Vector Machine + Adaboost	NA	Python	5
[27]	Qatar Biobank	987	Age, history of high cholesterol, history of diabetes, mother history of blood pressure, waist circumference, fruits & vegetables diet, physical activity, tobacco use, employment, education level, gender, age in years, nationality	5-fold cross-validation	Hypertensive: 14.3% Normal: 85.7%	Literature	Random Forest	0.869	WEKA	3

Table 2. Summary information of the selected articles. (Conclusion)

Article	Database	Sample Size	Features	Training and Validation Strategy	Balancing	Features Selection Strategy	Best Algorithm	AUC	System Used	Number of Citations (Scholar, until 15 Aug 2021)
[28]	Massachusetts University Amherst and National Health and Nutrition Survey	17,030	Pounds, age, sex, race, height, mean SBP, mean DBP, smoker, and cholesterol	66% - Train 33% - Test	Hypertensive: 19.9% Normal: 80.1%	Uninformed	Naive Bayes	NA	Python	2
[29]	Data collected by community health workers through door-to-door and camp-based screenings in the urban slums of Hyderabad, India	2,278	Left arm SBP, blood sugar, age, BMI, Left Arm DBP, weight, pulse rate, height, waist circumference, medication, urination, gender, diabetic family, HTN family, dizziness, smoking, numbness, tingling, dry tongue, heartache.	25 (iterative) 10-fold cross validation	Hypertensive: 26.4% Normal: 73.6%	Random Forest - Gini	Random Forest	0.792	Python	1
[30]	Korean National Health Insurance Corporation	4,707	BMI, DBP, total cholesterol and family history	Train and test split. Percentages were not informed.	Uninformed	Uninformed	Support Vector Machine	0.900	Non-Specified	1
[13]	Beijing We-Health Platform	8,253	BMI, age, weight, FPG, triglyceride, uric acid, hemoglobin, total cholesterol, urea, hematocrit, red blood cell, sex, white blood cell.	10-fold Cross-Validation 70% - train and 30% test	Hypertensive: 34.1% Normal: 65.9%	Random Forest	KNN	NA	Non-Specified	0
[31]	China Health and Nutrition Survey	3,015	26 nutritional features, age, and BMI	85% - train 15% - test	Hypertensive: 49.1% Negative: 50.9%	Literature	XGBoost	0.904	Python	0

Note: Body Mass Index (BMI), Metabolic Equivalents of Task (METs), Systolic Blood Pressure (SBP), Diastolic Blood Pressure (DBP), Heart Rate (HR), Hemoglobin A1c (HBA1C), Cardio-ankle Vascular Index (CAVI), High-density Lipid Cholesterol (HDL), Blood urea nitrogen (BUN), White blood cell (WBC), Red blood cell (RBC), Forced vital capacity (FVC), Predicted forced vital capacity predicted (FVCP), Forced expiratory volume in 1 s (FEV1), Predicted forced expiratory volume in 1 s predicted (FEV1P), ratio of forced expiratory volume in 1 s to forced vital capacity (FEV1FVC), Forced expiratory volume in 6 s (FEV6), Forced expiratory flow 25–75% (FEF25–75), Peak expiratory flow (PEF), Hypertension (HT) and Fasting Plasma Glucose (FPG).

DISCUSSION

The use of artificial intelligence to predict systemic arterial hypertension has the potential to improve targeted interventions and decrease the future incidence of the disease. To our knowledge, this is the first systematic review of machine learning studies for hypertension prediction. The studies presented an overall high Area Under the ROC Curves (AUROC), ranging from 0.766 to 1.00. The algorithms most frequently selected to perform the prediction were support vector machines, XGBoost and random forest.

An important challenge of machine learning for hypertension prediction is how to define the target variable. There are variations between guidelines, and, whenever possible, researchers should include clinical parameters associated with hypertension, such as risk factors, target organ damage, and imaging markers. HTN is traditionally defined as serial measurements of SBP greater than 140 mmHg or DBP equal to or greater than 90 mmHg [32]. However, studies demonstrate target-organ involvement with values above 115/75 mmHg, usually considered “low” [33]. Also, serological and clinical markers can be present before BP reaches values above 130-140/90 mm Hg, such as exaggerated responses to physical exercise, mild left ventricular hypertrophy, and the presence of microalbuminuria [34].

In order to facilitate clinical use by doctors, machine learning algorithms must be able to explain the mechanisms behind their decisions. While simpler models have higher interpretability, such as logistic regression and decision trees, they frequently have lower predictive performance than more rigorous algorithms such as decision tree ensembles like XGBoost and random forests. However, a growing number of techniques make these complex algorithms explainable, which is the case of the increasingly popular Shapley values, which uses coalitional game theory to identify the average contribution of a feature [35].

Although most studies identified by this review used open-source programming languages such as R and Python, most studies did not publicly make their prediction algorithms available. Due to its in-silico nature, machine learning experiments are often reproducible, and sharing the code can improve the applicability, technical advancements, and reproducible research.

Another critical concern in machine learning modeling is data leakage, considered one of the most frequent errors in data science [36]. Data leakage

can be defined as a training process that includes an outcome-informational feature, i.e., when a feature posteriorly associated with the outcome is included as a predictor. A clear example of data leakage in machine learning studies for hypertension is the inclusion of anti-hypertensive medication among the predictor variables. This scenario happened in one of the studies found by this systematic review, as previously identified [29].

Our study followed the PRISMA recommendations for conducting and reporting the results of systematic reviews. A funnel plot was not constructed due to the diversity of performance criteria. It was also not possible to conduct a meta-analysis due to the diversity of algorithms, methods, and outcomes among the individual studies. During the final stages of this study, a broad literature review was published on machine learning for hypertension [30]. However, it did not perform a systematic review, neither compared the results regarding the five elements of machine learning. Finally, most of the articles included are observational and did not assess, for example, whether there were clinical improvements for patients whose doctors had access to the results of the predictive algorithms.

CONCLUSION

This study aimed to identify the recent literature on machine learning for hypertension prediction. A total of 21 articles were selected, revealing a large diversity of dataset types, country origins, training strategies, hypertension definitions, feature selections, algorithms, and performance evaluation metrics. The scientific literature on machine learning for hypertension is rapidly improving, and there is a great potential for machine learning algorithms to improve preventive decisions and targeted policies for hypertension, but factors such as outcome definition, availability of the final code, predictive performance, explainability, and data leakage, need to be closely evaluated.

REFERENCES

Papers of particular interest, published recently, have been highlighted as: • Of importance •• Of major importance.

1. World Health Organization (2021) Hypertension. In: Overview. https://www.who.int/health-topics/hypertension#tab=tab_1. Accessed 9 Jun 2021

2. Carretero AO, Oparil S (2000) Clinical cardiology: New frontiers. Circulation 101:329–335

3. Messerli FH, Williams B, Ritz E (2007) Essential hypertension. Lancet 370:591–603

4. Manosroi W, Williams GH (2018) Genetics of Human Primary Hypertension: Focus on Hormonal Mechanisms. Endocr Rev. https://doi.org/10.1210/er.2018-00071

5. Onusko E (2003) Diagnosing secondary hypertension. Am Fam Physician 67:67–74

6. Charles L, Triscott J, Dobbs B (2017) AFP-secondary HTN- discoverting the underlying cause. Am Fam Physician 96:453–461

7. Cai L, Zhu Y (2015) The challenges of data quality and data quality assessment in the big data era. Data Sci J 14:1–10

8. Batista AF., Chiavegatto Filho ADP (2019) Machine Learning aplicado à Saúde. Workshop: Machine Learning. 19º Simpósio Bras. Comput. Apl. à Saúde. Soc. Bras. Comput.

9. van de Schoot R, de Bruin J, Schram R, et al (2021) An open source machine learning framework for efficient and transparent systematic reviews. Nat Mach Intell 3:125–133

10. Kumar V (2014) Feature Selection: A literature Review. Smart Comput Rev. https://doi.org/10.6029/smartcr.2014.03.007

11. Kwong EWY, Wu H, Pang GKH (2018) A prediction model of blood pressure for telemedicine. Health Informatics J 24:227–244

12. •• Sakr S, Elshawi R, Ahmed A, Qureshi WT, Brawner C, Keteyian S, Blaha MJ, Al-Mallah MH (2018) Using machine learning on cardiorespiratory fitness data for predicting hypertension: The Henry Ford exercise testing (FIT) Project. PLoS One. https://doi.org/10.1371/journal.pone.0195344

13. Ma Y, Yang B, Kang G, Hou B (2018) Hypertension Warning Model Based on Random Forest and Distance Metrics. In: 2018 IEEE Int. Conf. Bioinforma. Biomed. IEEE, pp 2274–2279

14. Heo BM, Ryu KH (2018) Prediction of prehypertenison and hypertension based on anthropometry, blood parameters, and spirometry. Int J Environ Res Public Health. https://doi.org/10.3390/ijerph15112571

15. Nour M, Polat K (2020) Automatic Classification of Hypertension Types Based on Personal Features by Machine Learning Algorithms. Math Probl Eng. https://doi.org/10.1155/2020/2742781

16. •• Ye C, Fu T, Hao S, et al (2018) Prediction of Incident Hypertension Within the Next Year: Prospective Study Using Statewide Electronic Health Records and Machine Learning. J Med Internet Res 20:e22

17. •• Ijaz MF, Alfian G, Syafrudin M, Rhee J (2018) Hybrid Prediction Model for type 2 diabetes and hypertension using DBSCAN-based outlier detection, Synthetic Minority Over Sampling Technique (SMOTE), and random forest. Appl Sci. https://doi.org/10.3390/app8081325

18. •• Nusinovici S, Tham YC, Chak Yan MY, Wei Ting DS, Li J, Sabanayagam C, Wong TY, Cheng CY (2020) Logistic regression was as good as machine learning for predicting major chronic diseases. J Clin Epidemiol 122:56–69

19. •• Pei Z, Liu J, Liu M, Zhou W, Yan P, Wen S, Chen Y (2018) Risk-Predicting Model for Incident of Essential Hypertension Based on Environmental and Genetic Factors with Support Vector Machine. Interdiscip Sci Comput Life Sci 10:126–130

20. Kanegae H, Suzuki K, Fukatani K, Ito T, Harada N, Kario K (2020) Highly precise risk prediction model for new-onset hypertension using artificial intelligence techniques. J Clin Hypertens 22:445–450

21. Soh DCK, Ng EYK, Jahmunah V, Oh SL, San TR, Acharya UR (2020) A computational intelligence tool for the detection of hypertension using empirical mode decomposition. Comput Biol Med. https://doi.org/10.1016/j.compbiomed.2020.103630

22. Xu F, Zhu J, Sun N, et al (2019) Development and validation of prediction models for hypertension risks in rural Chinese populations. J Glob Health. https://doi.org/10.7189/jogh.09.020601

23. Li C, Sun D, Liu J, Li M, Zhang B, Liu Y, Wang Z, Wen S, Zhou J (2019) A prediction model of essential hypertension based on genetic and environmental risk factors in northern han chinese. Int J Med Sci 16:793–799

24. Zhang L, Yuan M, An Z, et al (2020) Prediction of hypertension, hyperglycemia and dyslipidemia from retinal fundus photographs via deep learning: A cross-sectional study of chronic diseases in central China. PLoS One 15:1–11

25. López-Martínez F, Núñez-Valdez ER, Crespo RG, García-Díaz V (2020) An artificial neural network approach for predicting hypertension using NHANES data. Sci Rep 10:10620

26. Ambika M, Raghuraman G, SaiRamesh L (2020) Enhanced decision support system to predict and prevent hypertension using computational intelligence techniques. Soft Comput 24:13293–13304

27. AlKaabi LA, Ahmed LS, Al Attiyah MF, Abdel-Rahman ME (2020) Predicting hypertension using machine learning: Findings from Qatar Biobank Study. PLoS One 15:e0240370

28. Marin I, Goga N (2019) Hypertension Detection based on Machine Learning. In: Proc. 6th Conf. Eng. Comput. Based Syst. ACM, New York, NY, USA, pp 1–4

29. • Boutilier JJ, Chan TCY, Ranjan M, Deo S (2021) Risk Stratification for Early Detection of Diabetes and Hypertension in Resource-Limited Settings: Machine Learning Analysis. J Med Internet Res. https://doi.org/10.2196/20123

30. Patnaik R, Chandran M, Lee S-C, Gupta A, Kim C, Kim C (2018) Predicting the occurrence of essential hypertension using annual health records. In: 2018 Second Int. Conf. Adv. Electron. Comput. Commun. IEEE, pp 1–5

31. Liu Y, Li S, Jiang H, Wang J (2020) Exploring the relationship between hypertension and nutritional ingredients intake with machine learning. Healthc Technol Lett 7:103–108

32. Chobanian A V., Bakris GL, Black HR, et al (2003) Seventh report of the Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure. Hypertension 42:1206–1252

33. England TN (2001) Numb Er 18 of Cardiovascular Disease. English J 345:1291–1297

34. Giles TD, Berk BC, Black HR, Cohn JN, Kostis JB, Izzo Jr. JL, Weber MA (2005) Expanding the Definition and Classification of Hypertension. J Clin Hypertens 7:505–512

35. Kononenko I (2001) Machine learning for medical diagnosis: History, state of the art and perspective. Artif Intell Med 23:89–109

36. Nisbet R, Elder J, Miner GD (2009) Handbook of Statistical Analysis and Data Mining Applications. Academic Press

37. Amaratunga D, Cabrera J, Sargsyan D, Kostis JB, Zinonos S, Kostis WJ (2020) Uses and opportunities for machine learning in hypertension research. Int J Cardiol Hypertens 5:100027

Artigo 2: Machine learning for longitudinal mortality risk prediction in patients with malignant neoplasm in São Paulo, Brazil

ELSEVIER

Artificial Intelligence in the Life Sciences
Volume 3, December 2023, 100061

Research Article

Machine learning for longitudinal mortality risk prediction in patients with malignant neoplasm in São Paulo, Brazil

GFS Silva [a], LS Duarte [b], MM Shirassu [b], SV Peres [c], MA de Moraes [b], A Chiavegatto Filho [a]

Abstract

Artificial intelligence is becoming an important diagnostic and prognostic tool in recent years, as machine learning algorithms have been shown to improve clinical decision-making. These algorithms will have some of their most important applications in developing regions with restricted data collection, but their performance under this condition is still widely unknown. We analyzed longitudinal data from São Paulo, Brazil, to develop machine learning algorithms to predict the risk of death in patients with cancer. We tested different algorithms using nine separate model structures. Considering the area under the ROC curve (AUC-ROC), we obtained values of 0.946 for the general model, 0.945 for the model with the five main cancers, 0.899 for bronchial and lung cancer, 0.947 for breast cancer, 0.866 for stomach cancer, 0.872 for colon cancer, 0.923 for rectum cancer, 0.955 for prostate cancer, and 0.917 for uterine cervix cancer. Our results indicate the potential of building models for predicting mortality risk in cancer patients in developing regions using only routinely-collected data.

Keywords: machine learning, artificial intelligence, predictive model, cancer

(1) Corresponding author. E-mail: gabriel8.silva@usp.br

Introduction

Neoplasms are defined by abnormal tissue growth and can be classified as benign or malignant. Benign (noncancerous) neoplasms are characterized by slow and organized spread, the presence of well-defined borders, and the absence of an invasive character at both the tissue and organ levels. Malignant (cancerous) neoplasms, on the other hand, are characterized by often rapid and disorganized growth, with poorly defined borders and possible invasion of adjacent tissues and organs, implying the possibility of metastatic cancer [1, 2].

According to the World Health Organization, about 9.6 million people died of cancer worldwide in 2018, of which around 70% were in middle- and low-income countries [3]. In Brazil, according to the National Cancer Institute (INCA) [4], about 625,000 new cases were expected in 2020, based on estimates from before the SARS-coV-2 pandemic, and were distributed among cancers of the prostate, female breast, colon and rectum, trachea, bronchus and lung, and

stomach. As for the total number of deaths, according to the Brazilian Mortality Information System data (SIM), 224,829 deaths were caused by malignant neoplasms in 2020, and the most frequent were trachea, bronchus, and lung (28,516 deaths), breast (18,032 deaths), prostate (15,841 deaths), stomach (13,850 deaths), and colon (12,422 deaths) [5].

Artificial Intelligence (AI) has become a valuable tool in the field of medicine. Machine learning algorithms can identify patterns and trends from data that may not be readily apparent to the human eye. This allows medical professionals to make more accurate predictions about patient diagnosis and prognosis and make informed decisions about their treatment. Machine learning in healthcare has the potential to greatly improve patient outcomes and make the healthcare system more efficient.

Given the growing scenario of cancer cases in Brazil and around the world [4,6], it is increasingly important to improve prognostic decisions for cancer patients [7]. The aim of this work is to develop machine learning algorithms to predict the risk of death in cancer patients in order to provide inputs for their clinical management.

Material and Methods

Dataset Description

We analyzed data collected from the Hospital Cancer Registry (RHC) of the Oncocenter Foundation of São Paulo (FOSP/SP) [8], a public registry that monitors patients treated in the state of São Paulo since the year 2000. RHC has information on the cancer diagnosis, treatment, metastases, recurrences, age and sex of the patients, and data on the health facilities/organizations where the consultations were performed. The dataset includes a total of 99 variables and 1,085,380 patients from 2000 through September 2022.

All variables collected after the cancer diagnosis for each patient were removed. The algorithms were trained with twelve variables: sex, age, days between first physician visit and diagnosis, clinical stage of cancer, category of medical service, previous diagnosis, type of diagnosis, topography group, health region of residence [9], morphology, health institution habilitation, and health region of diagnosis. There are three levels to the variable category medical

service: 1) private care, 2) public care, 3) private care. The variable previous diagnosis presents binary information, 1 for patients who started longitudinal follow-up with a previous cancer diagnosis and 0 for patients without previous diagnosis. The variable type of diagnosis presents four categories: 1) clinical examination, 2) non-microscopic auxiliary resources, 3) microscopic confirmation and 4) no information. The variables cancer topography and cancer morphology are categorized with ICD-10 and ICD-O, respectively. The variables related to health region have seventeen distinct values, referring to the seventeen health regions in the state of São Paulo, Brazil. The variable health institution habilitation has fifteen categories: 1) High Complexity Oncology Care Unit (UNACON), 2) UNACON with Radiotherapy Service, 3) UNACON with Hematology Service, 5) Exclusive UNACON for Pediatric Oncology, 6) High Complexity Oncology Care Center (CACON), 7) CACON with Pediatric Oncology Service, 8) General Hospital with Oncological Surgery, 9) UNACON with Radiotherapy and Hematology Services, 10) UNACON with Radiotherapy, Hematology and Pediatric Oncology Services, 12) UNACON with Hematology and Pediatric Oncology Services, 13) Volunteer, 14) Inactive, 15) Exclusive UNACON for Pediatric Oncology with Radiotherapy Service. The full description of the dataset and its variables can be found in Supplementary Appendix B (Table B1).

Only patients with diagnoses from 2014 to 2017 were included, in order to avoid longer clinical effects after the diagnosis. Although the dataset included a small portion of population from other Brazilian states, we limited the algorithm development to residents of the state of São Paulo (93% of total patients). We included only patients with malignant neoplasms and excluded cases of non-melanoma of the skin as they had a low mortality rate. We analyzed adult patients regardless of sex. The final sample was composed of a total of 29,194 patients.

Outcome Definition

The original dataset contains four categories regarding the last available information about the patient: 1) alive without cancer, 2) alive with cancer, 3) death from cancer, and 4) death without further information. Our outcome of interest was patients with a confirmed cancer death between 12 and 24 months after the date of diagnosis. For the negative outcome, we included patients 1) alive without cancer or 2) alive with cancer between 12 and 36 months

after the date of diagnosis. Patients were removed if categorized as 4) death without other information.

Model Design

Considering the distinct cancer types, we tested different models to assess whether changing the strategy increased model performance. We first developed a general model for all cancer types. We then developed a model for the top five causes of cancer mortality (bronchus and lung, breast, stomach, colon, and rectum). We also trained specific models for the five most frequent causes and added two other models based on its growing importance for health vigilance: prostate cancer and cervix uteri cancer (in both cases the sex variable was not used as a predictor). We evaluated the models independently, without sharing any information during algorithm training. A summary of the model design is provided in Supplementary Appendix A (Figure A1).

Machine Learning Techniques

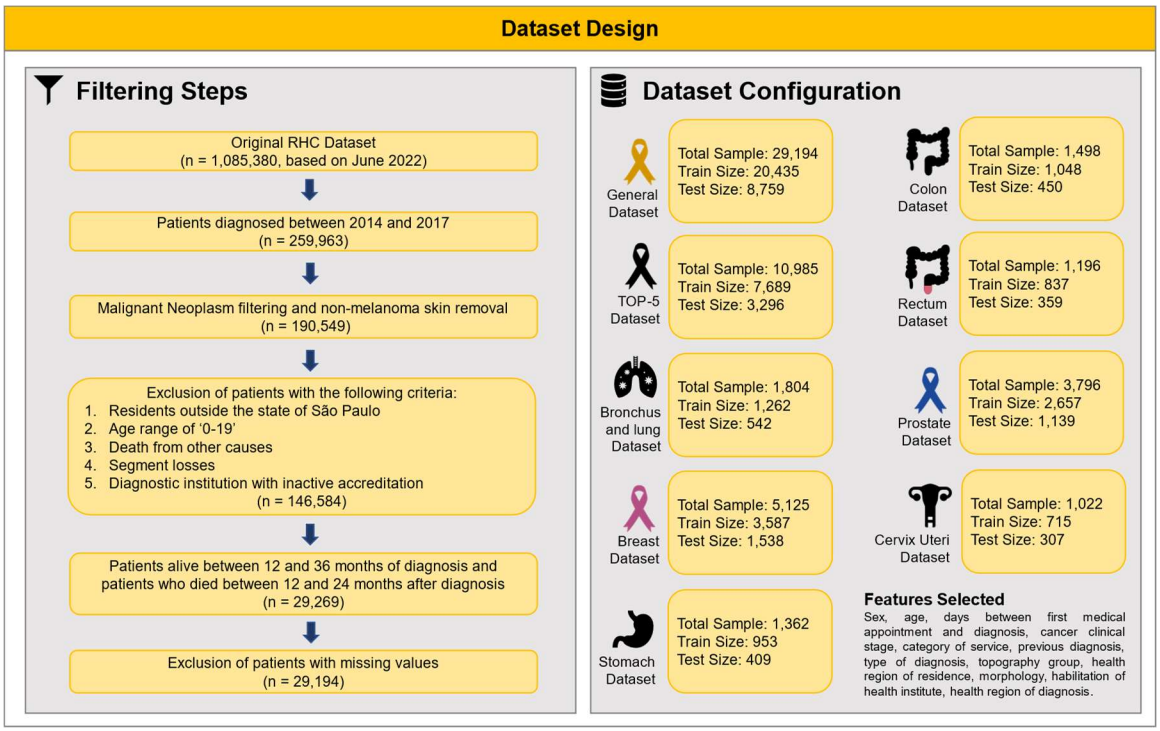
For quantitative variables, we performed normalization using the z-score (separately in training and test). For all qualitative variables, we separated each category using one-hot encoding. The variable type of diagnosis presented categorized missing (value 9) for twenty-four patients. We considered this a new category for the one hot encoding procedure. We also removed 75 patients due to the lack of any information in two variables: cancer stage (two patients) and difference in days between first medical appointment dates and diagnosis (73 patients).

We tested the predictive performance of six different machine learning algorithms: catboost [10], xgboost [11], lightgbm [12], gradient boosting classifier, random forest, logistic regression. For catboost, xgboost and lightgbm, we used their own Python packages. For the other algorithms, we used the scikit-learn library [13].

We used 10-fold cross-validation to select the hyperparameters in the training set with Hyperopt [14], which applies a Bayesian strategy for optimization, and RandomSearch. In the case of high-class imbalance (minority class representing under 25% of total outcomes), we applied the Synthetic Minority Oversampling Technique (SMOTE). Also in the training set, we applied the BORUTA [15] method for variable selection.

We then selected the best performing models from the training set (70% of the data) to evaluate their performance in the test set (30%). The complete structure of the datasets is presented in Figure 1.

Figure 1. Dataset filtering and configurations for predictive models' development.



The predictive performance of the models was evaluated on the test set using metrics such as area under the ROC curve (AUC-ROC), area under the precision-recall curve (AUC-PR), precision, recall, positive predicted value, negative predicted value, and F1-score. We also evaluated the performance of the algorithms in the 20% highest risk patients (20% k-tops), with metrics such as true positive, false positive, precision and recall. Finally, the interpretation and evaluation of the contribution of each variable to the outcome was obtained by calculating the Shapley values [16, 17, 18] for the test set. We followed the guidelines of the transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) [19].

Results and discussion

Descriptive Data Analysis

After data preprocessing, a total 29,194 patients were included in the study, most of whom were female (51.0%). A total of 27.7% of patients were

between 60-69 years old and 23.5% seventy or more. Regarding the clinical classification of the cancer, there was a relative balance between stages I (20.9%), II (20.3%), III (20.1%), IV (26.0%). The other categories accounted for about 12.6% of the total number of patients.

Regarding the category of health services, 72.7% of patients were diagnosed in public institutions, 26.6% in private services and 0.7% in individual private services. Of the total patients, 43.6% died between 12 and 24 months after the date of the malignant cancer diagnosis and 56.4% remained alive. The characteristics of the patients were similar in the training and testing data, as well as in the general dataset (Table 1). Additional summaries of cancer morphology and topography are provided in Tables B2 and B3 in the Supplementary Appendix B.

Table 1. Descriptive summary of full, train and test datasets.

Variable	Full Dataset	Death	Non-death	Train	Test
Sex					
Male	14,313 (49.0%)	7,027 (55.2%)	7,286 (44.2%)	9,972 (48.8%)	4,341 (49.6%)
Female	14,881 (51.0%)	5,697 (44.8%)	9,184 (55.8%)	10,463 (51.2%)	4,418 (50.4%)
Age					
20-29	929 (3.2%)	263 (2.1%)	666 (4.0%)	652 (3.2%)	277 (3.2%)
30-39	2,176 (7.5%)	636 (5.0%)	1,540 (9.4%)	1,534 (7.5%)	642 (7.3%)
40-49	3,862 (13.2%)	1,453 (11.4%)	2,409 (14.6%)	2,714 (13.3%)	1,148 (13.1%)
50-59	7,270 (24.9%)	3,234 (25.4%)	4,036 (24.5%)	5,077 (24.8%)	2,193 (25.0%)
60-69	8,093 (27.7%)	3,690 (29.0%)	4,403 (26.7%)	5,658 (27.7%)	2,435 (27.8%)
70+	6,864 (23.5%)	3,448 (27.1%)	3,416 (20.7%)	4,800 (23.5%)	2,064 (23.6%)
Clinical Stage					
I	6,107 (20.9%)	570 (4.5%)	5,537 (33.6%)	4,292 (21.0%)	1,815 (20.7%)
II	5,917 (20.3%)	1,371 (10.8%)	4,546 (27.6%)	4,119 (20.2%)	1,798 (20.5%)
III	5,876 (20.1%)	2,843 (22.3%)	3,033 (18.4%)	4,088 (20.0%)	1,788 (20.4%)
IV	7,602 (26.0%)	6,090 (47.9%)	1,544 (9.4%)	5,361 (26.2%)	2,241 (25.6%)
X	712 (2.4%)	414 (3.3%)	1,512 (9.2%)	482 (2.4%)	230 (2.6%)
Y	2,980 (10.2%)	1,436 (11.3%)	298 (1.8%)	2,093 (10.2%)	887 (10.1%)
Service Category					
Public	21,224 (72.7%)	11,865 (93.2%)	9,359 (56.8%)	14,840 (72.6%)	6,384 (72.9%)
Private	7,755 (26.6%)	803 (6.3%)	6,952 (42.2%)	5,448 (26.7%)	2,307 (26.3%)
Particular	215 (0.7%)	56 (0.4%)	159 (1.0%)	147 (0.7%)	68 (0.8%)

Outcome					
Death	12,724 (43.6%)	-	-	8,906 (43.6%)	3,818 (43.6%)
Non-death	16,470 (56.4%)	-	-	11,529 (56.4%)	4,941 (56.4%)

Model data structure

We developed a total of 42 machine learning algorithms considering nine root structures: 1) general model, 2) top 5 cause of death model, 3) bronchus and lung cancer model, 4) breast cancer model, 5) stomach cancer model, 6) colon cancer model, 7) rectum cancer model, 8) prostate cancer model, and 9) cervix uteri cancer model. Table 2 provides a descriptive summary of each model considering the total number of patients, total number of deaths and nondeaths, total mortality, and the number of patients in the training and test groups. There were notable imbalances according to the different models, with 29,194 patients in the general model and 1,022 in the cervix uteri cancer model, which may have affected the predictive performance of the algorithms.

Table 2. Description of the eight models developed for prediction of cancer mortality. (To be continued)

ID	Model	Variables	Total cases	Non-death	Death	Mortality rate	Train Size (70%)	Test Size (30%)
1	General	sex, age, medsvtodiag, cancerstage, servicecat, prevdiag, diagbase, topogroup, rras, morpho, habilit, rrasofserv	29,194	16,470	12,724	43.6%	20,435	8.759
2	Top-5 cause of death	sex, age, medsvtodiag, cancerstage, servicecat, prevdiag, diagbase, topogroup, rras, morpho, habilit, rrasofserv	10,985	5,937	5,048	46.0%	7,689	3.296
3	Bronchus and lung	sex, age, medsvtodiag, cancerstage, servicecat, prevdiag, diagbase, rras, morpho, habilit, rrasofserv	1,804	468	1,336	74.1%	1,262	542
4	Breast	sex, age, medsvtodiag, cancerstage, servicecat, prevdiag, diagbase, rras, morpho, habilit, rrasofserv	5,125	3,845	1,280	25.0%	3,587	1.538

Table 2. Description of the eight models developed for prediction of cancer mortality. (Conclusion)

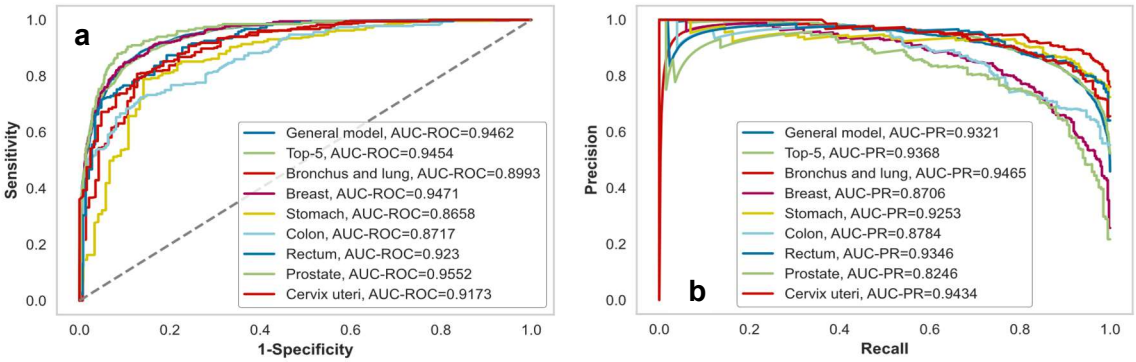
ID	Model	Variables	Total cases	Non-death	Death	Mortality rate	Train Size (70%)	Test Size (30%)
5	Stomach	sex, age, medsvtodiag, cancerstage, servicecat, prevdiag, diagbase, rras, morpho, habilit, rrasofserv	1,362	401	961	70.6%	953	409
6	Colon	sex, age, medsvtodiag, cancerstage, servicecat, prevdiag, diagbase, rras, morpho, habilit, rrasofserv	1,498	740	758	50.6%	1,048	450
7	Rectum	sex, age, medsvtodiag, cancerstage, servicecat, prevdiag, diagbase, rras, morpho, habilit, rrasofserv	1,196	483	713	59.6%	837	359
8	Prostate	age, medsvtodiag, cancerstage, servicecat, prevdiag, diagbase, rras, morpho, habilit, rrasofserv	3,796	3,232	664	17.5%	2,657	1.139
9	Cervix uteri	age, medsvtodiag, cancerstage, servicecat, prevdiag, diagbase, rras, morpho, habilit, rrasofserv	1,022	416	609	59.6%	715	307

medsvtodiag: difference in days between first medical appointment dates and diagnosis, **cancerstage**: cancer clinical stage, **servicecat**: category of service, **prevdiag**: previous diagnosis, **diagbase**: type of diagnosis, **topogroup**: cancer topography, **rras**: regional net of healthcare (residence), **morpho**: cancer morphology, **habilit**: qualification of the health establishment, **rrasofserv**: regional net of healthcare (service).

Algorithms Performance

The catboost algorithm presented the best performance on all models except stomach, where Gradient Boosting performed better in terms of AUC-ROC. Figure 2 presents the performance of the best prediction algorithms for each of the models considering the AUC-ROC test set criterion (a) and AUC-PR (b). We found good discrimination performance for the nine models. All models presented AUC-ROC values of at least 0.871 and six of them reached values above 0.900. The general model presented the best overall performance, with AUC-ROC of 0.946 and AUC-PR of 0.932. The model for the five main causes of death (top-5) presented an AUC-ROC of 0.945 and an AUC-PR of 0.937.

Figure 2. Predictive performance of best algorithm for each model regarding AUC-ROC (a) and AUC-PR (b).



In the general model, we tested combinations of hyperparameter optimization (Hyperopt and RandomSearch) and variable selection (BORUTA). However, the best predictive performance was achieved for the raw model, with AUC-ROC of 0.946, recall 0.855, specificity 0.889, precision 0.857, F1-score 0.856, and AUC-PR 0.932 (Table 3). Although changes in algorithm settings improved at least one scoring metric (i.e., the general model with RandomSearch had a recall of 0.858), the raw model was the one that presented the best AUC-ROC. When BORUTA was used, there was a significant reduction in the number of predictors (497 to 93) without a large loss in predictive performance (AUC-ROC of 0.946 for the raw model versus 0.945 for the model with BORUTA and without hyperparameter optimization). A similar pattern to the general model was observed for the top 5 causes of death model. Complete results for all training strategies can be found in Supplementary Appendix B (Table B4). The hyperparameters of each model are also available in Supplementary Appendix B (Tables B5, B6, B7, B8, B9, B10, B11, B12, B13).

Table 3. Predictive performance of best algorithm for each model.

ID	Model	Best Algorithm	Hypermeter Tunning	Feature Selection	Resample	Accuracy	AUC-ROC	Recall	Specifi city	Prec.	F1	AUC-PR
1	General	CatBoost Classifier	None	None	None	0.8743	0.9462	0.8549	0.8893	0.8565	0.8557	0.9321
2	Top-5 cause of death	CatBoost Classifier	None	None	None	0.8686	0.9454	0.8581	0.8776	0.8564	0.8572	0.9368
3	Bronchus and lung	CatBoost Classifier	None	None	SMOTE	0.8561	0.8993	0.9152	0.6879	0.8929	0.9039	0.9465
4	Breast	CatBoost Classifier	None	None	None	0.8973	0.9471	0.7214	0.9558	0.8445	0.7781	0.8706
5	Stomach	Gradient Boosting	RandomSearch	None	None	0.8093	0.8658	0.9343	0.5083	0.8207	0.8738	0.9253
6	Colon	CatBoost Classifier	None	None	None	0.7578	0.8717	0.7763	0.7387	0.7532	0.7646	0.8784
7	Rectum	CatBoost Classifier	Hyperopt	None	None	0.8412	0.9230	0.9159	0.7310	0.8340	0.8731	0.9346
8	Prostate	CatBoost Classifier	None	None	None	0.9210	0.9552	0.7487	0.9574	0.7884	0.7680	0.8246
9	Cervix uteri	CatBoost Classifier	Hyperopt	None	None	0.8306	0.9173	0.9235	0.6935	0.8164	0.8667	0.9434

We also evaluated the performance of the algorithms in the top 20% (20% k-tops) of patients with the highest mortality risk (Table 4). The general model had 1,749 patients in the group, of whom 1,703 died, giving the algorithm a precision of 97,37% and a recall of 100% in this high-risk group. For the top 5 causes of death model, 659 individuals were in the 20% highest risk patients, of which 645 died, resulting in a precision of 97.88% and a recall of 100%.

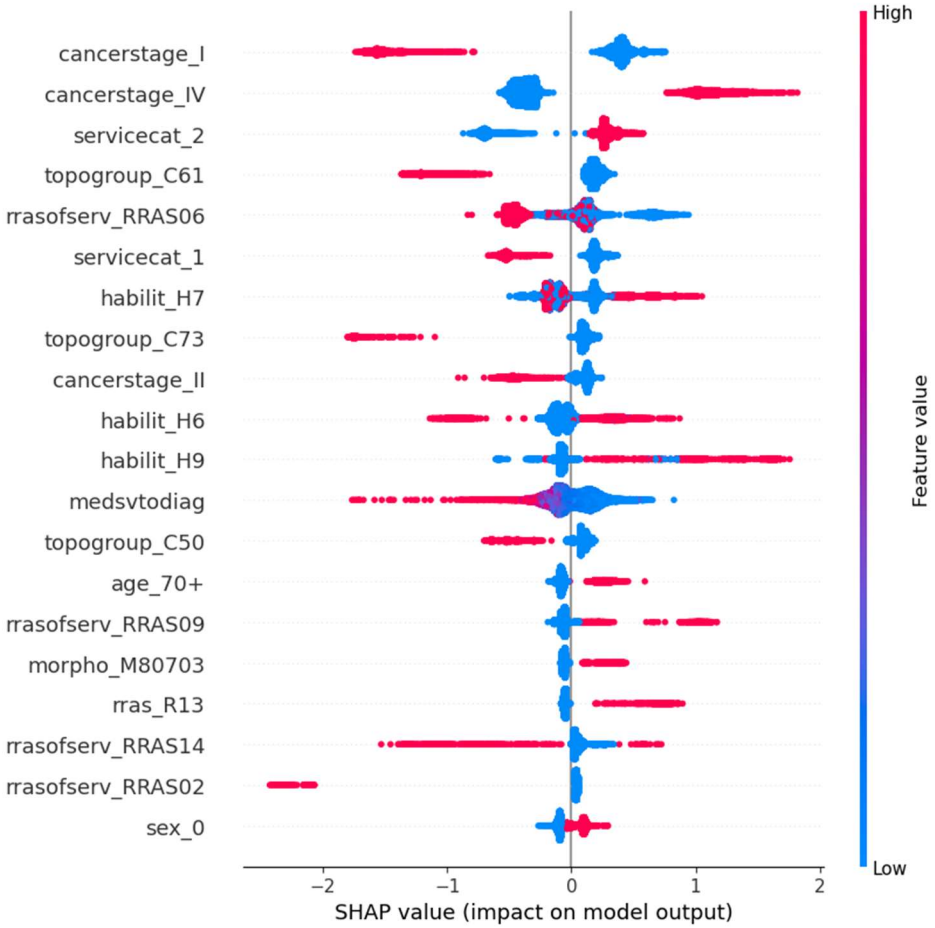
Table 4. Predictive performance of best algorithm for each model based on 20% individuals with the highest risk of death.

ID	Model	Total Patients	Real Positive	Positive Prediction	True Positive	False Positive	Precision	Recall
1	General	1,749	1,703	1,749	1,703	46	0.9737	1.0000
2	Top-5 cause of death	659	645	659	645	14	0.9788	1.0000
3	Bronchus and lung	109	107	109	107	2	0.9817	1.0000
4	Breast	308	263	308	263	45	0.8539	1.0000
5	Stomach	82	78	82	78	4	0.9512	1.0000
6	Colon	90	88	90	88	2	0.9778	1.0000
7	Rectum	72	70	72	70	2	0.9722	1.0000
8	Prostate	228	166	189	149	40	0.7884	0.8976
9	Cervix uteri	62	60	62	60	2	0.9677	1.0000

Model Interpretation

In order to interpret the decision-making process of the algorithms, we calculated the Shapley values. In the general model (Figure 3), the cancer stage during the first diagnosis was the most important predictor. Stage I patients were more likely to be classified negatively (non-death), whereas stage IV patients were more significant for the positive outcome (death). The variable on the category of service provided was also important for the outcome. Category 2 (public service) increased mortality prediction, whereas category 1 (private service) showed a greater propensity for patient survival. The other main predictive variables refer to the topography of cancer, regional net of healthcare service (rrasofserv) and regional net of healthcare service (rras). The plots of Shapley values for the other models can be found in Supplementary Appendix A (Figures A2, A4, A6, A8, A10, A12, A14, A16).

Figure 3. Top twenty predictors of risk of death from cancer 12 to 24 months after diagnosis. General Model, with Catboost Classifier.

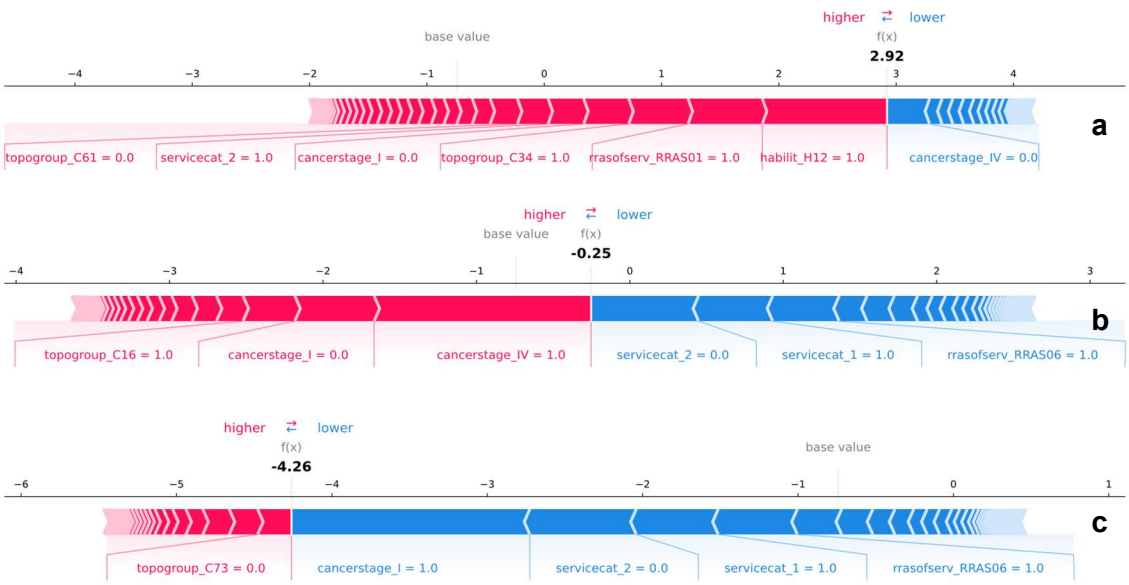


cancerstage_I: cancer stage I, **cancerstage_IV**: cancer stage IV, **servicecat_2**: public care service, **topogroup_C61**: cancer topography ICD C-61 (malignant neoplasm of prostate), **rrasofserv_RRAS06**: regional net of healthcare (service) 06, **servicecat_1**: private care service, **habilit_H7**: qualification H7 CACON with Pediatric Oncology Service, **topogroup_C73**: cancer topography ICD C-73 (malignant neoplasm of thyroid gland), **cancerstage_II**: cancer stage II, **habilit_H6**: qualification H6 CACON, **habilit_H9**: qualification H9 9 - UNACON with Radiotherapy and Hematology Services, **medsvtodiag**: difference in days between first medical appointment dates and diagnosis, **topogroup_C50**: cancer topography ICD C-50 (malignant neoplasm of breast), **age_70+**: age group of 70 years or more), **rrasofserv_RRAS09**: regional net of healthcare (service) 09, **morpho_M80703**: cancer morphology 80703 (squamous cell carcinoma, NOS), **rras_R13**: regional net of healthcare (residence) 13, **rrasofserv_RRAS14**: regional net of healthcare (service) 14, **rrasofserv_RRAS02**: regional net of healthcare (service) 02.

We also randomly selected three patients (high risk, medium risk, and low risk) to highlight the individual interpretation of results (Figure 4). The first (a) was a true positive (risk of 0.972) with the expected Shapley value was 2.92. The variables that contributed for the prediction of positive outcome were public care service (servicecat_2 = 1), cancer stage different from I (cancerstage_I = 0), cancer topography ICD C-34 (malignant neoplasm of bronchus and lung),

regional net of service 01 (rrasofserv_RRAS01 = 1) and the qualification of the healthcare institution (code 12, UNACON with Hematology and Pediatric Oncology Services). A second patient (b) classified as a false negative was selected. The total risk score was 0.4782, which led the algorithm to classify the patient incorrectly as alive during the period. We observed that there was balance in the aggregate of the contribution of the predictors, highlighting the importance of cancer stage IV to increase Shapley value and the non-public health service to decrease it. For patient c, a true negative classified as low risk, the most important characteristic to a low expected Shapley value were cancer stage I and non-public health service. Visualizations of the individual Shapley values for the other models are available in Supplement A (figures A3, A5, A7, A9, A11, A13, A15, A17).

Figure 4. Main predictors of risk of death from cancer between 12 and 24 months after diagnosis for three randomly selected individuals: a) high risk of death (true positive with 0.972 score), b) medium risk of death (false negative with 0.478 score) and c) low risk of death (true negative with 0.017 score), general model with Catboost Classifier.



Patient a) topogroup_C61: cancer topography ICD C-61 (malignant neoplasm of prostate), **servicecat_2:** public care service, **cancer_stage1:** cancer stage I, **topogroup_C34:** cancer topography ICD C-34 (malignant neoplasm of bronchus and lung), **rrasofserv_RRAS01:** regional net of healthcare (service) 01, **habilit_H12:** qualification H12 UNACON with Hematology and Pediatric Oncology Services, **cancerstage_IV:** cancer stage I. **Patient b) servicecat_1:**

private care service, **morpho_80703:** cancer morphology 80703 (squamous cell carcinoma, NOS), **cancerstage_I:** cancer stage I, **cancerstage_IV:** cancer stage IV, **servicecat_2:** public private care service, **rrasofserv_RRAS06:** regional net of healthcare (service) 06. **Patient c) topogroup_C73:** cancer topography ICD C-73 (malignant neoplasm of thyroid gland), **cancerstage_I:** cancer stage I, **servicecat_2:** public care service, **servicecat_1:** private care service, **rrasofserv_RRAS06:** regional net of healthcare (service) 06. Zero value are interpreted as the absence of the characteristic and one as the presence.

General vs specific models

To understand the best strategy regarding the types of models (general or specific for each cancer), we performed a comparison between the performance of the general algorithm for all cancers versus the specific algorithms for bronchus and lung, breast, stomach, colon, rectum, prostate, bronchus and lung, colon, and uterine cervix (Table 5). Based on the AUC-ROC, the general model performed better in bronchus and long, stomach and colon. For breast, rectum, prostate, and cervix uteri the model performed better for the specific case. For bronchus and lung cancer, the area under the curve increased from 0.899 (model specific) to 0.927 (general model) and the precision from 0.893 to 0.907. For stomach cancer, there was an increase in AUC-ROC (0.866 to 0.926), precision (0.821 to 0.924). This scenario was repeated for colon cancer (AUC-ROC 0.753 to 0.848).

Table 5. Comparison of the predictive performance between the specific algorithms for each type of cancer and the general algorithm.

Cancer Type	Model	Test size	Real Positive	True Positive	False Positive	True Negative	False Negative	Precision	Recall	AUC-ROC
Bronchus and Lung	General	486	367	340	35	84	27	0.9067	0.9264	0.9265
	Specific	542	401	367	44	97	34	0.8929	0.9152	0.8993
Breast	General	1192	206	146	29	957	60	0.8343	0.7087	0.9460
	Specific	1538	384	277	51	1103	107	0.8445	0.7214	0.9471
Stomach	General	383	277	258	26	80	19	0.9085	0.9314	0.9255
	Specific	409	289	270	59	61	19	0.8207	0.9343	0.8658
Colon	General	468	241	200	36	191	41	0.8475	0.8299	0.9241
	Colon	450	228	177	58	164	51	0.7532	0.7763	0.8717
Rectum	General	319	197	177	26	96	20	0.8719	0.8985	0.9163
	Colon	359	214	196	39	106	18	0.8340	0.9159	0.9230
Prostate	General	1192	206	146	29	957	60	0.8343	0.7087	0.9460
	Specific	1139	199	149	40	900	50	0.7884	0.7487	0.9552
Cervix Uteri	General	326	215	177	21	90	38	0.8939	0.8233	0.8850
	Specific	307	183	170	42	82	13	0.8164	0.9235	0.9173

Discussion

We found that all models achieved an AUC-ROC higher than 0.86 to predict cancer mortality using only routinely-collected data. Our results also indicated that a general algorithm, that included all cancer mortality, performed in most cases better than cancer-specific algorithms.

Information about mortality risk after cancer diagnosis can be an important input to support clinical decisions. These algorithms can be integrated into mobile devices, electronic medical records, or online resources, to help doctors in making more informed decisions about treatment options and to allocate healthcare resources more effectively.

We obtained a high predictive performance without the use of omics or image data, which is a promising result in the field of oncology especially in low-income regions. We were able to develop an approach to compare the use of a general model with a model for the main causes of death, and with models specific to each type of cancer. Most of the studies developed in the literature are specific for a given type of cancer, due to the selected data sets, using either from image data or through structured data [20, 21]. The use of data from a cancer registry allowed for achieving consistent results in all proposed models, while at the same time providing a real-world dataset, with recent cases and different types of cancer.

The study has a few limitations. First, the algorithms were developed with data from São Paulo, Brazil, so there should be caution in transferring its conclusions to other contexts. Second, considering the filtered sample, we had 1,391 follow-up losses that could have disproportionally altered the results. Third, we excluded patients younger than 20 years due to the presence of different biological mechanisms that lead to cancer deaths for this group, so the results refer only to adult patients.

Conclusion

In conclusion, the nine final models developed for predicting risk of death in cancer patients presented high predictive performance. The algorithms can be an important tool to help prioritize treatment decisions and patient allocation in cancer treatments, especially in low-income regions. Future work should explore

the proposed methodological structure and evaluate its predictive performance in new settings with different routinely collected data.

Ethical statement

This work was evaluated and approved by the Research Ethics Committee of the Faculty of Public Health of the University of São Paulo (CAAE: 65375722.9.0000.5421)

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data and code availability

The main results of this research were published in this article and in supplementary appendix A and B. The RHC/FOSP data are publicly available in https://www.fosp.saude.sp.gov.br/fosp/diretoria-adjunta-de-informacao-e-epidemiologia/rhc-registro-hospitalar-de-cancer/. The code developed for predictive modeling can be obtained upon request.

Acknowledgment

We would like to acknowledge the São Paulo State Health Department, Special Health Fund for Mass Immunization and Disease Control (FESIMA/SES/SP) for supporting and funding this research.

Funding

This work was supported by the São Paulo State Health Department, Special Health Fund for Mass Immunization and Disease Control (FESIMA/SES/SP).

Supplementary Material

Supplementary material for this article is available as an online appendix[2].

[2] https://github.com/gabriel1710/tese_material_suplementar/tree/master/Artigo%202

References

[1] Patel, A. (2020). Benign vs Malignant Tumors. JAMA Oncology, 6(9):1488–1488.
[2] Thuler, L. C. S., Sant'Ana, D. R., and Rezende, M. C. R. (2011). Abc do câncer: abordagens básicas para o controle do câncer. In ABC do câncer: abordagens para o controle do câncer, pages 127–127.

[3] World Health Organization (2022). Fact sheets: cancer.

[4] INCA (2019). Estimativa 2020: incidência de câncer no Brasil.

[5] Ministério da Saúde, Brasil. (2022). Sistema de informação sobre mortalidade (SIM).

[6] Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., and Jemal, A. (2018). Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians, 68(6):394–424. [1]

[7] Iqbal, M. J., Javed, Z., Sadia, H., Qureshi, I. A., Irshad, A., Ahmed, R., Malik, K., Raza, S., Abbas, A., Pezzani, R., et al. (2021). Clinical applications of artificial intelligence and machine learning in cancer diagnosis: looking into the future. Cancer cell international, 21(1):1–11.

[8] Secretaria de Estado da Saúde. (2022). Fundação Oncocentro de São Paulo. Registro Hospitalar de Câncer: banco de dados. São Paulo, Brasil.

[9] Lavras, C. (2011). Atenção primária à saúde e a organização de redes regionais de atenção à saúde no brasil. Saúde e Sociedade, 20:867–874.

[10] Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: unbiased boosting with categorical features. Advances in neural information processing systems, 31.

[11] Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). New York, NY, USA: ACM. https://doi.org/10.1145/2939672.2939785

[12] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., … Liu, T.-Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems, 30, 3146–3154.

[13] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. the Journal of machine Learning research, 12, 2825-2830.

[14] Bergstra, J., Yamins, D., Cox, D. D. (2013) Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures. To appear in Proc. of the 30th International Conference on Machine Learning (ICML 2013).

[15] Kursa MB, Rudnicki WR (2010). “Feature Selection with the Boruta Package.” Journal of Statistical Software, 36(11), 1–13. https://doi.org/10.18637/jss.v036.i11

[16] Lundberg, S., and Lee, S.-I. A unified approach to interpreting model predictions. In NIPS (2017).

[17] Lundberg, S.M., Nair, B., Vavilala, M.S. et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. Nat Biomed Eng 2, 749–760 (2018). https://doi.org/10.1038/s41551-018-0304-0

[18] Lundberg, S.M., Erion, G., Chen, H. et al. From local explanations to global understanding with explainable AI for trees. Nat Mach Intell 2, 56–67 (2020). https://doi.org/10.1038/s42256-019-0138-9

[19] Moons, K. G. M. et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): Explanation and elaboration. Ann. Intern. Med. 162(1), W1-73 (2015).

[20] Sharma, A., Rani, R. A Systematic Review of Applications of Machine Learning in Cancer Prediction and Diagnosis. Arch Computat Methods Eng 28, 4875–4896 (2021). https://doi.org/10.1007/s11831-021-09556-z

[21] Kumar, Y., Gupta, S., Singla, R. et al. A Systematic Review of Artificial Intelligence Techniques in Cancer Prediction and Diagnosis. Arch Computat Methods Eng 29, 2043–2070 (2022). https://doi.org/10.1007/s11831-021-09648-w

Artigo 3: Unsupervised machine learning for the regionalization of healthcare management in noncommunicable diseases

Gabriel Ferreira dos Santos Silva
School of Public Health, University of São Paulo, Brazil
Luciane Simões Duarte
São Paulo State Health Department, Division of Noncommunicable diseases, São Paulo, Brazil
Mirian Matsura Shirassu
São Paulo State Health Department, Division of Noncommunicable diseases, São Paulo, Brazil
Marco Antonio de Moraes
São Paulo State Health Department, Division of Noncommunicable diseases, São Paulo, Brazil
Alexandre Dias Porto Chiavegatto Filho
School of Public Health, University of São Paulo, Brazil

ABSTRACT

Objectives: Noncommunicable Diseases (NCD) are complex and highly prevalent throughout the world. Machine learning (ML) algorithms can improve the organization of healthcare management by identifying clusters of areas with similar challenges regarding NCD. The aim of this study was to develop a regional clustering method for NCD morbidity and mortality profile in the most populous State of Brazil (São Paulo) using ML.

Methods: We developed clusters for the 645 municipalities of São Paulo State (SSP), Brazil, using mortality and hospital morbidity variables from 2010-2019. Coefficients were age-standardized, and exploratory spatial analyses were performed with LISA clusters. Hierarchical clustering, k-means, and SKATER were used to create the final clusters with varying numbers of groups (k).

Results: LISA cluster identified regions of high or low mortality and hospital morbidity in the SSP. The SKATER algorithm presented a better consistence for regionalization purposes with 17 clusters (k = 17). Radar charts were used to analyze the coefficients in each cluster by specific disease group of NCD and identified the presence of high mortality due to diabetes in five different clusters, and high values for cardiovascular disease mortality in one cluster. We then developed a final interactive online application, in which it is possible to interact with all 17 clusters identified by the study.

Conclusions: Unsupervised ML algorithms were able to group areas with similar morbidity and mortality profile for NCD. The use of clustering methods for identifying epidemiologically similar municipalities can be an important tool for planning and managing healthcare systems.

Keywords: Noncommunicable Diseases; Artificial Intelligence; Machine Learning; Cluster Analysis; Disease Hotspot; Community Networks.

INTRODUCTION

Noncommunicable Diseases (NCD) are caused by a combination of genetic, physiological, environmental, and behavioral factors and have a long-lasting latency period (1). The World Health Organization (WHO) highlights four main NCD - cardiovascular disease, malignant neoplasms, chronic respiratory diseases, and diabetes - as its focus of global action (2-3). NCD account for approximately 74% of deaths worldwide (1), with the four main groups responsible for 57% of NCD deaths (4). In the state of São Paulo (SSP), the most populous of Brazil, 56.8% of deaths result from at least one of the four main NCDs (5).

In Brazil, previous studies have shown that the installment of Health Care Networks (RAS) are the most effective response to the fragmentation of the Brazilian public healthcare system (6), and in 2012 the SSP was divided into 17 Regional Health Care Networks (RHCN) (7). Although the RHCN was based on overall factors such as demographic, geographic, socioeconomic, epidemiological and health structures, several changes in population, epidemiological profiles and clinical medicine have occurred since its implementation.

In this context, machine learning (ML) algorithms, a branch of artificial intelligence that supports decision-making trough classification, regression, or clustering tasks, can bring significant benefits to improve health policy organization and management. Therefore, the aim of this study was to develop regional clustering methods for NCD morbidity and mortality in SSP using machine learning techniques.

METHODS

Data sources

We used a total of four public domain databases. First, individual mortality data and cause of death records were obtained from the Brazilian Mortality Information System (SIM) (5). We also collected individual records from Hospital Information System of the Unified Health System (SIH/SUS) (8), and current population estimative from the Ministry of Health/SVS/DASNT/CGIAE (9). Finally, we also obtained health insurance beneficiary data from the National Supplementary Health Agency (ANS) (10). We analyzed data separately for all

the 645 municipalities of SSP from 2010 until 2019 in order to avoid the effects of the COVID-19 pandemic on the morbidity and mortality profile of the population.

Variables of the study

Regional clusters for specific disease groups of the main NCD were developed, according to the International Classification of Diseases - 10th revision (ICD-10): diabetes (E10-E14), chronic respiratory diseases (J30-J97, except J36), diseases of the circulatory system (I00-I99), and malignant neoplasms (C00-C97). The morbidity and mortality coefficients were age-standardized using the direct method, considering the following age groups: 0 years, 1-9 years, 10-19 years, 20-39 years, 40-59 years, 60-79 years, and 80 years or more. The standard population used was the total estimate of the SSP, according to the Brazilian Ministry of Health (9). Mortality coefficients were calculated for 100,000 residents. Hospital morbidity in the Brazilian Unified Health System (SUS) was calculated for 10,000 SUS-dependent residents. This specific population group was obtained from the ANS database (10).

Exploratory data analysis

Individual data from the SIM and SIH/SUS were first grouped by municipalities and year of occurrence, along with their respective population estimates. Morbidity and mortality coefficients were developed for each year of the study. To evaluate the totality of the study area, which comprised all 645 municipalities, the sum of deaths, hospitalizations, and populations of each municipality was calculated from 2010 to 2019. This sum mathematically equals the consolidated average coefficients weighted by the populations for this decade. The coefficients were normalized using the z-score in order to group the coefficients (mortality and hospital morbidity in SUS). A hard-stop of three standard deviations up and own was applied to control individual outlier clusters.

We performed a descriptive analysis of the coefficients, presenting measures of central tendency (mean and median) and dispersion (variance, standard deviation, maximum value, minimum value, and quartiles). The coefficient of variation (CV) was also calculated to evaluate the relative dispersion of data in relation to its mean.

Spatial exploratory analysis

Spatial autocorrelation analysis is commonly used to investigate patterns and clusters of diseases. To identify neighborhoods with low or high spatial correlation, we used the Local Indicator of Spatial Association (LISA) cluster (11), which evaluates spatial autocorrelation based on the variables and geographic restrictions. A 60-kilometer arc length was adopted as the neighborhood criterion, considering the maximum distance between the centroids of two neighboring municipalities. The objective of this initial exploratory analysis was to investigate the structure of disease patterns and identify their spatial dispersions of high and low prevalence, based on its average value.

Clustering Methods

We used unsupervised learning algorithms to identify regional groups in the SSP based on mortality and hospital morbidity data for each of the 645 municipalities. Three clustering methods were tested: k-means (12), hierarchical clustering (13), and Spatial 'K'luster Analysis by Tree Edge Removal (SKATER) (14). The selection of the number of clusters (k) was performed using the NBClust library (15), which tests the value of k based on 30 distinct indices by varying all combinations of number of clusters, distance measures, and clustering methods. The final clusters were created with k=17. To ensure contiguity, the queen neighborhood criterion of order one was considered for the construction of the weight matrix. However, this generated a spatial outlier (“Ilhabela”), which was kept as an outlier for evaluation purposes. For the construction of the cluster map, we forced the outlier grouping to the closest municipalities. The clusters were developed with R and Geoda software (16).

RESULTS

Descriptive and exploratory analysis

Table 1 presents the consolidated means of age-adjusted mortality and hospital morbidity coefficients in the SUS for NCDs, from 2010 to 2019. Regarding mortality coefficients, cardiovascular diseases ranked first with 187.70 deaths/100,000 inhabitants, followed by malignant neoplasms with 107.61 deaths/100,000 inhabitants, chronic respiratory diseases with 35.59

deaths/100,000 inhabitants, and diabetes with 28.11 deaths/100,000 inhabitants. The mortality coefficient for diabetes had very high variability, ranging from 6.47 to 68.89 with a coefficient of variation of 35.13%. For the consolidated means of hospital morbidity coefficients, a similar behavior was observed, with the same rank of diseases. Diabetes presented the lowest value, with 9.46 hospitalizations/10,000 SUS-dependent inhabitants and the highest coefficient of variation (CV) of 72.65%.

Table 1. Descriptive summary of mortality and hospital morbidity coefficients in the Brazilian Unified Health System (SUS) for Noncommunicable Diseases (NCD) in the state of São Paulo, between the years 2010 and 2019.

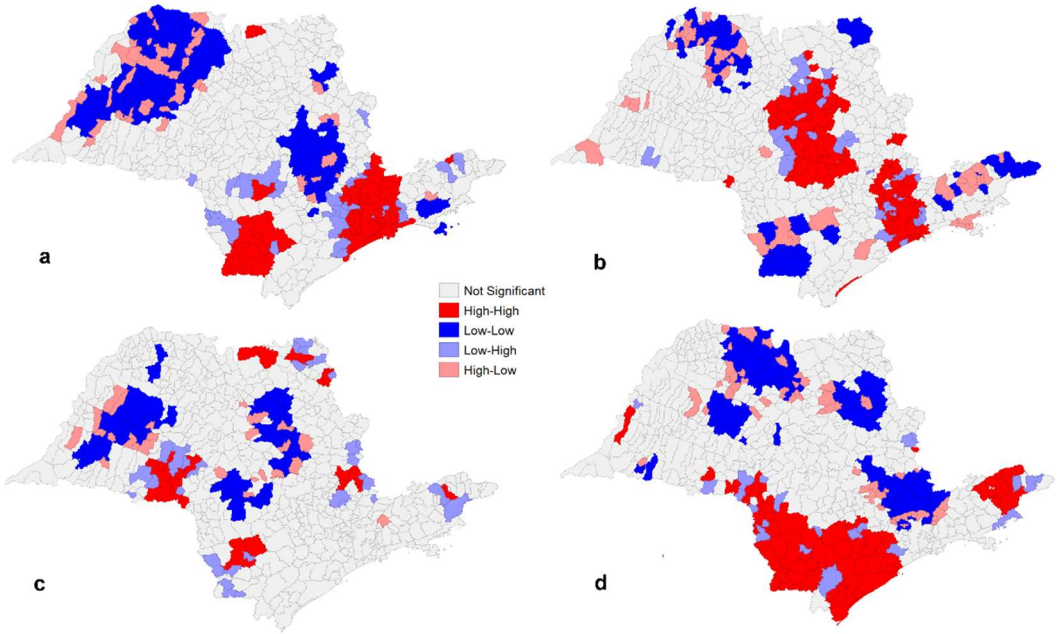
Coefficient[1]	Mean	Median	Variance	Standard Deviation	CV (%)	1st quartile	3rd quartile	Minimum	Maximum
Circulatory System Mortality	184.70	180.80	1 205.56	34.72	18.80%	160.00	202.80	111.40	406.70
Malignant Neoplasms Mortality	107.61	108.40	210.68	14.51	13.49%	98.36	116.89	59.27	166.30
Chronic Respiratory Mortality	38.59	37.95	74.91	8.65	22.43%	33.20	43.13	15.54	81.95
Diabetes Mortality	28.11	26.74	97.49	9.87	35.13%	21.10	33.84	6.47	68.89
Circulatory System Hospital Morbidity	99.77	95.62	1 349.93	36.74	36.83%	75.47	117.32	32.29	335.29
Malignant Neoplasms Hospital Morbidity	44.43	43.02	209.56	14.48	32.58%	34.18	52.84	8.29	120.42
Chronic Respiratory Hospital Morbidity	34.50	30.60	306.50	17.51	50.74%	23.08	42.11	8.08	150.02
Diabetes Hospital Morbidity	9.46	7.91	47.24	6.87	72.65%	5.26	11.79	0.99	86.33

(1) Mortality indicators were calculated based on a population of 100,000 inhabitants. Hospital morbidity indicators in the Brazilian Unified Health System (SUS) refer to hospitalizations in SUS. Their population bases are 10,000 SUS-dependent inhabitants. All indicators were age-adjusted.

Spatial exploratory analysis

LISA clusters were developed for mortality and hospital morbidity coefficients in SUS for NCD (Figure 1). No clear visual pattern between cardiovascular diseases (a), malignant neoplasm (b), chronic respiratory diseases (c), and diabetes (d) was identified (Figure 1). However, the southern region of the state stands out with a significant spatial correlation for high values (High-high) for diabetes, indicating a region with statistically superior mortality coefficients to the rest of the SSP. For cardiovascular diseases (a) and malignant neoplasms (b), a high correlation in the Metropolitan Region of São Paulo (MRSP), in the eastern region of the state, was observed. For chronic respiratory diseases (c), there was no clear pattern of regionalization.

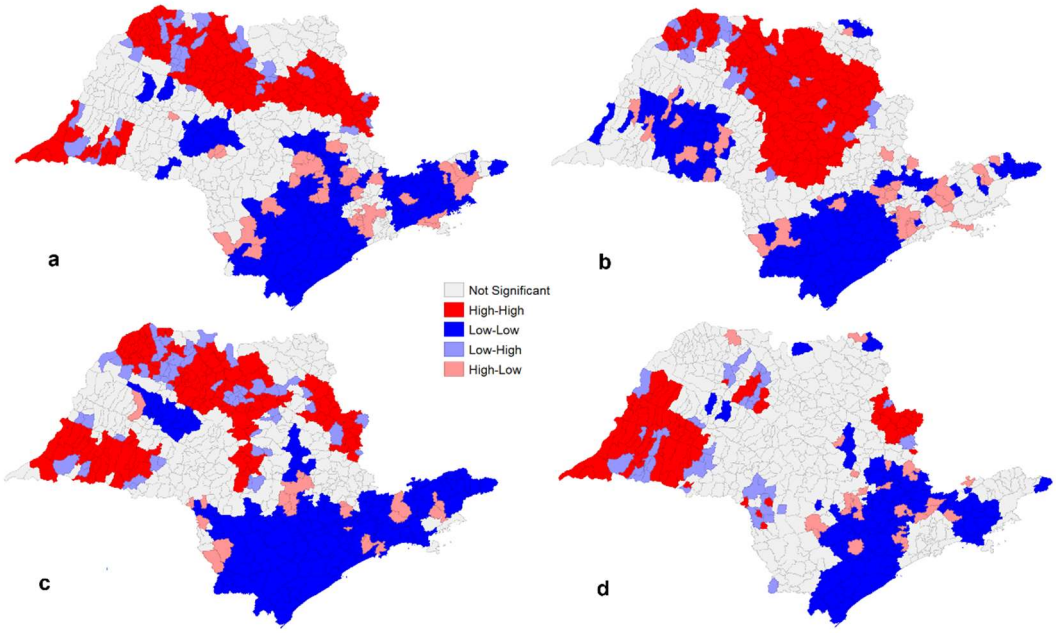
Figure 1. LISA Cluster of mortality coefficients for NCD (a) cardiovascular diseases, (b) malignant neoplasms, (c) chronic respiratory diseases, and (d) diabetes in the SSP, between 2010 and 2019.



Regarding hospital morbidity in SUS, Figure 2 shows that the southern and southeastern regions of the state presented high spatial correlation in the four disease groups, but for lower values (Low-low), indicating groups of municipalities with hospital morbidity lower than the SSP average. This fact is especially important considering that some of these municipalities are in high-high areas for mortality and low-low for hospital morbidity in SUS, i.e. they are

regions that grouped together due to high mortality for some NCD and, simultaneously, low hospital morbidity in relation to the average values of the SSP.

Figure 2. LISA Cluster of hospital morbidity coefficients for NCD (a) cardiovascular diseases, (b) malignant neoplasms, (c) chronic respiratory diseases, and (d) diabetes in the SSP, between 2010 and 2019.



Regional clustering by specific disease group of NCD

Determination of k values

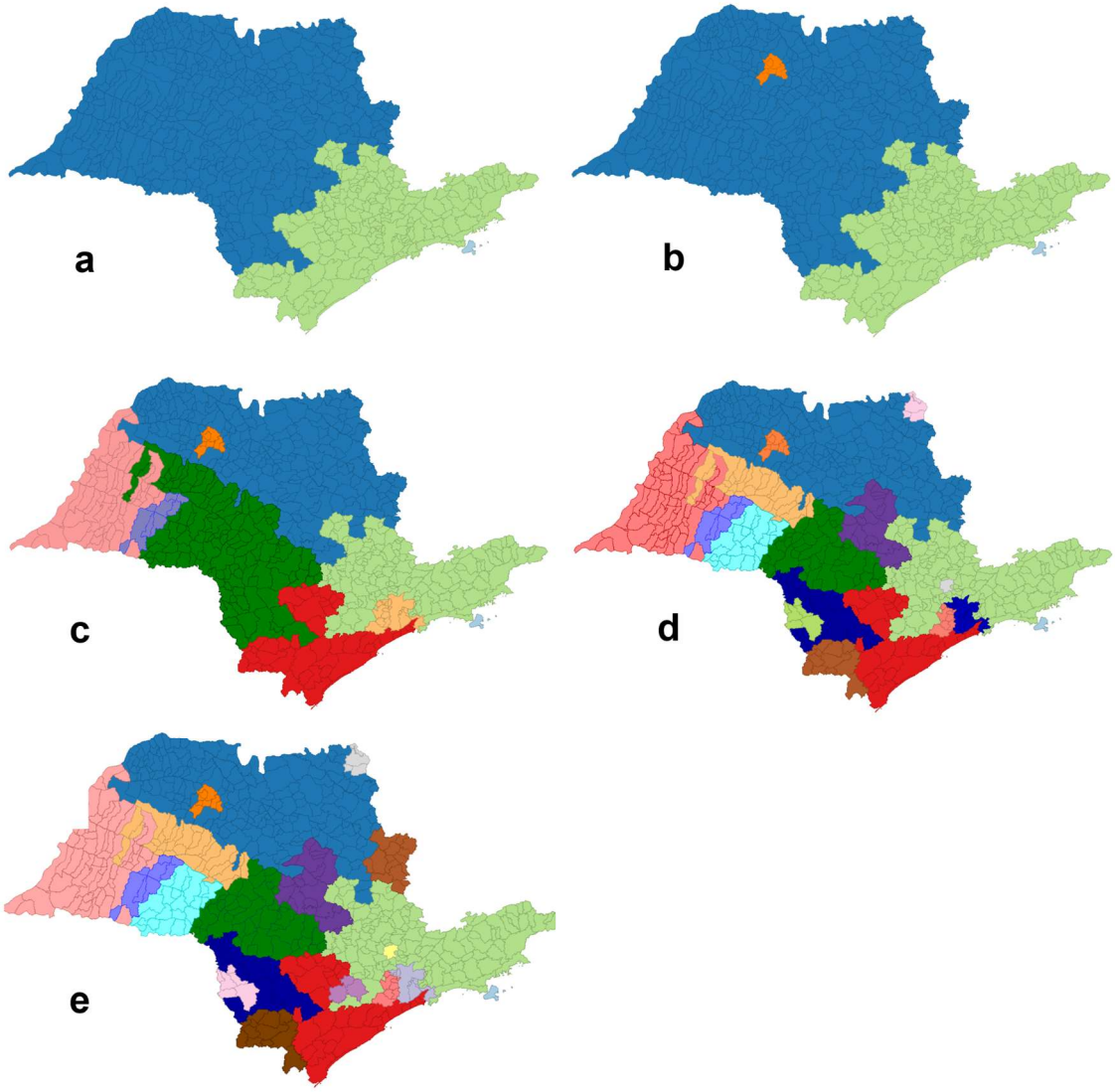
Clustering algorithms are considered an area of unsupervised machine learning. However, supervision is required to delimit the number of clusters (k) in order to split the data. Using the methods developed through the NBClust library in R, 7 algorithms proposed 2 clusters, 8 algorithms proposed 3 clusters, 1 algorithm proposed 5 clusters, 2 algorithms proposed 8 clusters, 1 algorithm proposed 12 clusters, 3 algorithms proposed 19 clusters and 1 algorithm proposed 20 clusters. The proposals to develop 5, 12, and 20 clusters were excluded as they were suggested by only 1 algorithm. Clusters with k equal to 2, 3, 8, and 19 clusters were tested, in addition to k equal to 17 clusters, as this is the current number of RHCN in SSP.

Clustering methods

A total of three unsupervised methods were applied to identify morbimortality clusters: k-means, hierarchical clustering, and SKATER. The k-means and hierarchical clustering methods identified relevant clusters, however, difficulties were observed in regionalization, as municipalities were not grouped contiguously into visible clusters, even with the inclusion of their latitude and longitude as grouping variables. Thus, the SKATER method was chosen for the development of regional clusters by specific group of NCD.

Figure 3 presents the SKATER cluster by specific disease group of NCD in the SSP. It can be observed that for k=2 (a), a northwest-southeast division was established. Starting from k=3 (b), a small group consisting of the municipalities of Jaci, Poloni, Planalto, União Paulista, Monte Aprazível, Neves Paulista, and Nipoã was identified with higher values for some coefficients in relation to the means of the other clusters (mortality rates due to chronic respiratory diseases, hospital morbidity in SUS for diabetes, diseases of the circulatory and chronic respiratory systems). Using k=17 (d), there was the presence of an outlier (Jundiaí), even after the application of the hard-stop technique, which presented high coefficients of mortality due to malignant neoplasms, hospital morbidity in SUS due to chronic respiratory diseases and hospital morbidity in SUS due to malignant neoplasms in relation to the circumscribed municipalities. For k=19, no considerable changes were observed. The cluster with k=17 was then selected, as it represents the number of current RHCN in SSP.

Figure 3. Regional SKATER cluster by specific disease group of NCD in the SSP, between the years 2010 to 2019, with (a) k = 2, (b) k = 3, (c) k = 8, (d) k = 17, and (e) k = 19.

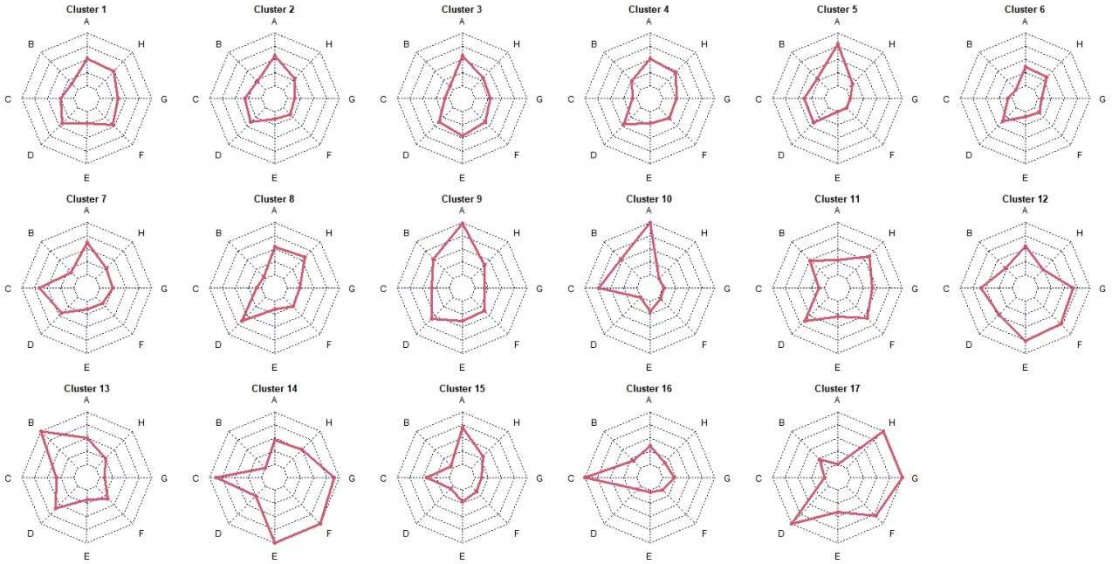


Clusters composed of the variables: age-adjusted general mortality rate for diabetes for the period 2010 to 2019, age-adjusted general mortality rate for circulatory system diseases for the period 2010 to 2019, age-adjusted general mortality rate for chronic respiratory diseases for the period 2010 to 2019, age-adjusted general mortality rate for malignant neoplasms for the period 2010 to 2019, age-adjusted hospital morbidity rate for diabetes in the Brazilian Unified Health System (SUS) for the period 2010 to 2019, age-adjusted hospital morbidity rate for circulatory system diseases in SUS for the period 2010 to 2019, age-adjusted hospital morbidity rate for chronic respiratory diseases in SUS for the period 2010 to 2019, age-adjusted hospital morbidity rate for malignant neoplasms in SUS for the period 2010 to 2019.

Radar charts were developed to analyze the coefficients (mortality and hospital morbidity in the SUS) in each cluster by specific disease group of NCD (Figure 5). High values were observed for mortality due to diabetes in clusters 5, 7, 9, 10, and 15, and high values for mortality due to cardiovascular diseases in cluster 13. For cardiovascular diseases, cluster 13 exhibited the highest evident

value. Regarding chronic respiratory diseases, mortality rates were notable in clusters 7, 10, 14, and 16. High mortality values were observed in clusters 8, 9, 13, and 17 for malignant neoplasms. In an overall perspective, clusters 14 and 17 displayed a greater imbalance between hospital mortalities and morbidities.

Figure 5. Radar charts for mortality and hospital morbidity coefficients in the Brazilian Unified Health System (SUS) by specific disease group of NCD, in the period from 2010 to 2019, for the SSP, normalized via z-score. Average values per cluster.



A: Age-adjusted mortality rate for diabetes mellitus per 100,000 inhabitants, B: Age-adjusted mortality rate for cardiovascular diseases per 100,000 inhabitants, C: Age-adjusted mortality rate for chronic respiratory diseases per 100,000 inhabitants, D: Age-adjusted mortality rate for malignant neoplasms per 100,000 inhabitants, E: Age-adjusted hospitalization morbidity rate for diabetes mellitus per 10,000 SUS-dependent inhabitants, F: Age-adjusted hospitalization morbidity rate for cardiovascular diseases per 10,000 SUS-dependent inhabitants, G: Age-adjusted hospitalization morbidity rate for chronic respiratory diseases per 10,000 SUS-dependent inhabitants, H: Age-adjusted hospitalization morbidity rate for malignant neoplasms per 10,000 SUS-dependent inhabitants.

Based on these results, an interactive online application was developed, in which it is possible to consult the raw values for each of the municipalities[3]. The results regarding general causes showed a similar behavior. Clusters were tested with k equal to 2, 3, 18, 19, and 20, and clusters proposed by only one algorithm were excluded. Additionally, SKATER cluster with k equal to 17 was tested[4].

[3] https://gabriel1710.github.io/clusters_ses/map_ncd_causes_cluster.
[4] https://gabriel1710.github.io/clusters_ses/map_general_causes_cluster.

DISCUSSION

We used unsupervised machine learning to perform regional clustering for NCD morbidity and mortality. The results identified regions with high and low mortality and hospital morbidity in the public healthcare system. For the regionalization of SSP municipalities using the unsupervised SKATER algorithm, clusters with similar mortality and hospital morbidity had geographic proximity, allowing for the local segmentation of healthcare regions with similar challenges.

RHCN is an important tool for the care of patients with NCD (6), which is increasingly necessary in the face of accelerated demographic transition (17) and epidemiological transition of triple disease burden, especially regarding chronic conditions (6). RHCN decreases the fragmentation of the system, which is fundamental for the care of NCD (6).

The healthcare surveillance attention model in Brazil is oriented towards identifying health risks, the likelihood of groups developing a disease or presenting a health condition, and determining damages, i.e. quantifying deaths, sequelae, or cases of diseases and health conditions (18). Although it was conceived from a collective perspective, since data analysis is done at the population level, its interventions can be directed at the individual level through preventive measures (18). In this regard, artificial intelligence is a promising surveillance technology for health analysis, especially spatial clustering.

Spatial clustering methodologies have been examined in several studies to identify disease clusters and guide further investigation. Torabi and Rosychuk (2011) (19) analyzed childhood cancer clusters in Alberta, Canada, using five popular methods and found potential clusters in the south-central part of the province. Rajabi et al. (2018) (20) explored spatial patterns of cardiovascular disease in Sweden, identifying hotspots in northern Sweden and clusters in central Sweden. Ramis et al. (2015) (21) conducted a case-control study on childhood cancer in Spain, observing variations in spatial distribution but no statistically significant clusters. These studies have provided important methodological foundations for the development of this analysis. Building upon these findings and insights, we further contributed to the field of spatial epidemiology by investigating the main regional chronic disease clusters for morbidity and mortality, and their potential implications for public health interventions.

This study has a few limitations. First, we used population estimates for calculating the mortality coefficients, resulting in fluctuations compared to the officially census number (which were last collected in 2010). Similarly, the SUS-dependent population, used for calculating hospital morbidity in the public sector, also comes from official local estimates. Another potential limitation is that mortality and hospital morbidity in SUS due to ill-defined causes have a relevant participation in the total composition of their respective indicators, corresponding to about 5% and 2% of total cases, respectively. Lastly, the use of geographic constraints to assess local healthcare management can limit the performance of cluster analysis since the separation between groups is not solely based on the nature of the variables.

CONCLUSION

Our study identified 17 regional clusters to support local healthcare management for health surveillance and promotion. The use of unsupervised machine learning algorithms is a promising tool to improve the efficiency of local healthcare management especially in developing regions with large budget constraints.

ETHICAL STATMENT

This work was evaluated and approved by the Research Ethics Committee of the Faculty of Public Health of the University of São Paulo (CAAE: 65,375,722.9.0000.5421)

FUNDING

This work was supported by the São Paulo State Health Department, Special Health Fund for Mass Immunization and Disease Control (FESIMA/SES/SP).

DECLARATION OF COMPETING INTEREST

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

ACKNOWLEDGEMENT

We would like to acknowledge the São Paulo State Health Department, Special Health Fund for Mass Immunization and Disease Control (FESIMA/SES/SP) for supporting and funding this research.

REFERENCES

1. World Health Organization. Noncommunicable Diseases. Key facts. World Health Organization, 2022. Available from: https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases. Accessed: March 31, 2023.
2. United Nations General Assembly. Political declaration of the high-level meeting of the General Assembly on the prevention and control of non-communicable diseases. New York: United Nations, 2011. Available from: https://digitallibrary.un.org/record/710899/?ln=en. Accessed: March 31, 2023.
3. United Nations General Assembly. Political declaration of the high-level meeting of the General Assembly on the prevention and control of non-communicable diseases. New York: United Nations, 2018. Available from: https://digitallibrary.un.org/record/1648984. Accessed: March 31, 2023.
4. World Health Organization. Noncommunicable diseases country profiles 2018. World Health Organization, 2018. Available from: https://apps.who.int/iris/handle/10665/274512. Accessed: March 31, 2023.
5. Brasil. Ministério da Saúde. Mortalidade São Paulo. TabNet: indicadores de saúde. Brasília: Ministério da Saúde, Brasil, 2023a. Available from: http://tabnet.datasus.gov.br/cgi/deftohtm.exe?sim/cnv/obt10sp.def. Accessed: March 31, 2023.
6. Mendes EV. As redes de atenção à saúde. Brasília: Organização Pan-Americana da Saúde, 2011.
7. Conselho de Secretários Municipais de São Paulo (COSEMS/SP). São Paulo, 2011. Available from: https://saude.sp.gov.br/resources/ses/perfil/gestor/homepage/redes-regionais-de-atencao-a-saude-no-estado-de-sao-paulo/redes-regionais-de-atencao-a-saude-rras/termo_de_referencia_redes_regionais.pdf. Accessed: March 31, 2023.

8. Brasil. Ministério da Saúde. Morbidade hospitalar do SUS - por local de residência - São Paulo. TabNet: indicadores de saúde. Brasília: Ministério da Saúde, Brasil, 2023b. Available from: http://tabnet.datasus.gov.br/cgi/deftohtm.exe?sim/cnv/obt10sp.def. Accessed: March 31, 2023.
9. Brasil. Ministério da Saúde. População residente - estudo de estimativas populacionais por município, idade e sexo 2000-2021. Brasília: Ministério da Saúde, Brasil, 2023c. Available from: http://tabnet.datasus.gov.br/cgi/deftohtm.exe?ibge/cnv/popsvsbr.def. Accessed: March 31, 2023.
10. Agência Nacional de Saúde (ANS). Dados e Indicadores do Setor. Beneficiários de planos privados de saúde. Beneficiários por municípios. 2023. Available from: http://www.ans.gov.br/anstabnet/cgi-bin/dh?dados/tabnet_02.def. Accessed: March 31, 2023.
11. Anselin L. Local indicators of spatial association—LISA. Geographical analysis. 1995;27(2):93-115.
12. Lloyd SP. Least squares quantization in PCM. IEEE Transactions on Information Theory. 1982;28(2):129-137.
13. Murtagh F, Contreras P. Algorithms for hierarchical clustering: an overview. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 2012;2(1):86-97.
14. Lage JP, Assunção RM, Reis EA. A minimal spanning tree algorithm applied to spatial cluster analysis. Electronic Notes in Discrete Mathematics. 2001;7:162-165.
15. Charrad M, Ghazzali N, Boiteau V, Nicknafs A. NbClust: Determining the Best Number of Clusters in a Data Set. R package version 3.0.1. Available from: https://cran.r-project.org/web/packages/NbClust/NbClust.pdf. Accessed Feb 25, 2022.
16. Anselin L, Syabri I, Kho Y. GeoDa: an introduction to spatial data analysis. In: Fischer MM, Getis A (eds.), Handbook of Applied Spatial Analysis. Springer, Berlin, Heidelberg; 2010. p. 73-89.
17. United Nations. World Population Prospects 2019, Volume I: Comprehensive Tables. New York: United Nations; 2019. Available from:

https://population.un.org/wpp/publications/Files/WPP2019_Volume-I_Comprehensive-Tables.pdf. Accessed Mar 31, 2023.

18. Paim JA. Modelos de atenção à Saúde no Brasil. In: Giovanella L, Escorel S, Lobato LVC, Noronha JC, Carvalho AI, organizers. Políticas e Sistema de Saúde no Brasil. 2ª ed. Rio de Janeiro: FIOCRUZ; 2012.
19. Torabi M, Rosychuk RJ. An examination of five spatial disease clustering methodologies for the identification of childhood cancer clusters in Alberta, Canada. Spatial and spatio-temporal epidemiology. 2011;2(4):321-330.
20. Rajabi M, Mansourian A, Pilesjö P, Åström DO, Cederin K, Sundquist K. Exploring spatial patterns of cardiovascular disease in Sweden between 2000 and 2010. Scandinavian journal of public health. 2018;46(6):647-658.
21. Ramis R, Gomez-Barroso D, Tamayo I, Garcia-Perez J, Morales A, Pardo Romaguera E, et al. Spatial analysis of childhood cancer: a case/control study. PLoS One. 2015;10(5):e0127273.

Artigo 4: Improving Health Management of Chronic Non-Communicable Diseases: Machine Learning Algorithms for Surveillance and Performance Evaluation in Healthcare Systems

Gabriel Ferreira dos Santos Silva
School of Public Health, University of São Paulo, Brazil
Luciane Simões Duarte
São Paulo State Health Department, Division of Non-communicable diseases, São Paulo, Brazil
Mirian Matsura Shirassu
São Paulo State Health Department, Division of Non-communicable diseases, São Paulo, Brazil
Marco Antônio de Moraes
São Paulo State Health Department, Division of Non-communicable diseases, São Paulo, Brazil
Alexandre Chiavegatto Filho
School of Public Health, University of São Paulo, Brazil

Abstract

Chronic non-communicable diseases (NCD) are a significant burden on healthcare systems worldwide. In order to improve management and surveillance of NCD, machine learning (ML) algorithms can provide important insights regarding performance evaluations. The objective of this study was to identify municipalities with higher or lower-than-expected mortality rates due to NCD in order to provide insights into the factors contributing to these outcomes. A comprehensive dataset was built, comprising demographic, socioeconomic, and infrastructure variables collected from the most populous state in Latin America (São Paulo, Brazil). ML algorithms, including lasso regressor, ridge regressor, elastic net, extra tree regressor, lightGBM, catboost, and xgboost were employed to develop predictive models for age-adjusted premature NCD mortality (AAPNM). A nested cross-validation approach was used for model training and evaluation, considering different population filters based on municipality size. The catboost regressor achieved the best performance for the municipalities with 50,000 inhabitants and over, with a RMSE of 37.246, MSE of 1387.228, and R^2 of 0.528, based on 9 predictors selected by the Boruta technique. Additional exploratory analyses revealed significant differences between municipalities with AAPNM above, below, and within the expected range, especially for variables such as average household income and unemployment rate. The results demonstrate the potential of using machine learning algorithms to identify municipalities requiring attention on NCD mortality.

Keywords: chronic non-communicable diseases, machine learning, predictive modeling, surveillance, performance evaluation, healthcare systems

INTRODUCTION

Non-communicable diseases (NCD), such as cardiovascular diseases, cancer, chronic respiratory diseases and diabetes, account for a significant proportion of deaths worldwide, posing a major challenge to healthcare systems (WHO, 2022). The proper management of these conditions requires continuous surveillance and careful assessment of health systems to ensure that patients receive the best possible care, especially in developing regions. This analysis

can be used to identify gaps in care provision, evaluate the effectiveness of prevention and treatment interventions, identify best practices and approaches, and inform health policy decisions.

Mortality from NCD, especially when premature (i.e. in population aged between 30 and 60 years), can be a relevant indicator for assessing health management, as it reflects not only the disease prevalence but also the quality and effectiveness of healthcare provided to patients (WHO, 2023). Low premature mortality rates indicate that prevention and treatment interventions are functioning adequately, and that health management is successfully controlling and treating the conditions effectively. On the other hand, a high mortality rate may indicate challenges in health management, such as issues with early disease detection, lack of access to appropriate care, and/or inadequacy of treatment interventions. Therefore, mortality from NCD can be an important indicator for healthcare managers to identify areas for improvement, in order to develop strategies to enhance care quality and reduce mortality (Budreviciute et al., 2020; Ministério da Saúde, 2014).

The use of machine learning (ML) techniques for surveillance and assessment of health management in NCD has become a promising new area, as these techniques enable automated and rapid analysis of large amounts of data, identifying patterns and correlations that would be difficult for the human eye (Davenport & Kalakota, 2019). This allows for potentially improving early case identification, enhancing risk prediction, and increasing the effectiveness of health interventions.

The objective of this study is to first develop ML algorithms capable of calculating the expected mortality rates for NCD in the most populous state of Latin America, i.e. São Paulo (SSP), Brazil, and then to identify its municipalities with observed mortality rates that deviate from this expected value.

METHODS

The SSP, located in the Southeast region of Brazil, is the most economically developed in the country according to the Brazilian Institute of Geography and Statistics (IBGE) (IBGE, 2023). With an estimated population of approximately 41 million inhabitants as of the 2010 census, São Paulo is home

to nearly a quarter of the country's total population, and approximately 11 million reside in the state capital, the city of São Paulo.

Regarding healthcare, the SSP boasts a relatively well-developed healthcare infrastructure, featuring renowned and references hospitals and medical research centers (SES/SP, 2023). However, like many parts of Brazil, the public healthcare system faces challenges, such as the need to improve access to healthcare in less privileged areas (Travassos et al., 2006). The city of São Paulo and other large municipalities, such as Campinas and São José dos Campos, are also affected by air pollution, heavy traffic, and the typical challenges of urban areas in ensuring quality of life (Dapper et al., 2016; Zerbini et al., 2009).

The SSP is divided into 645 municipalities, spread across 17 health regions. Despite being the country's largest economic hub, inequality is evident in the lives of residents and in the indicators presented in Table 1A from the Supplementary Appendix A (table_methods.xlsx), extracted from the last complete Brazilian census (2010). The average income per household for all the 645 municipalities of the state was 700.254 according to the 2010 census. However, expressive variability is observed in the quartile analysis, with minimum values of 308.690 and maximum values of 2008.980, indicating that there are municipalities with average income per household 6.5 times higher than others. A similar scenario occurs for other indicators, as inequality becomes visible in terms of selective waste collection, GDP per capita, the population living on less than half a minimum wage, as well as the Gini index.

Data Source

Three public data sources were used: 1) the last Brazilian census with full available data, i.e. the 2010 Census conducted by the Brazilian Institute of Geography and Statistics (IBGE), 2) the Ministry of Social Development (MDS), and 3) the Mortality Information System (SIM).

Outcome Definition

The outcome of the prediction models is the age-adjusted premature NCD mortality (AAPNM) between 2010 and 2019. Mortality data were derived from individual death records obtained through SIM. For the NCD group, the following

main causes of death from the International Classification of Diseases - 10th revision (ICD-10) of the World Health Organization (WHO, 2010) were considered: diabetes (E10-E14), chronic respiratory diseases (J30-J97, excluding J36), diseases of the circulatory system (I00-I99), and malignant neoplasms (C00-C97). Premature NCD mortality rates were calculated using the following formula:

$$Premature\ NCD\ Mortality = 100{,}000 \times \frac{\sum NCD\ death_{\ 2010\ to\ 2019}}{10 \times (Premature\ Populati_{\ \ 2010})}$$

Direct age-adjustment using age groups 30-39, 40-49, 50-59, and 60-69, was performed to mitigate the age effects within the premature group, which has a considerable range and heterogeneity of mortality. The official census population count of SSP from 2010 was used as the standard population for the age-adjustment.

Predictors

We used demographic, social, infrastructure, and economic variables as predictors. Variables directly related to healthcare management were not included in the models, in order to assess the expected mortality based on characteristics not amenable to direct healthcare policies. The description of the predictors is available in Supplementary Table 2A (table_methods.xlsx).

Outliers

To prevent performance issues and considering that the municipalities face challenges in managing chronic diseases, an outlier analysis was performed. The upper limit was calculated as the third quartile plus 1.5 times the interquartile range. The lower limit was obtained by subtracting 1.5 times the interquartile range from the first quartile. Municipalities that had AAPNM above the upper limit were considered as upper outliers. Similarly, municipalities with mortality values below the lower limit were considered as lower outliers. Although the positive outliers were excluded from the training process, they were properly identified and labeled as points of attention for further analysis, due to their high AAPNM in comparison to the other municipalities.

Machine Learning techniques

Data was preprocessed following current best practices for predictive modeling in ML (Lones, 2021). Variables with a missing value percentage equal to or greater than 40% were removed from the study. For variable pairs with a correlation above 0.90, the one with the lowest correlation with the outcome was excluded. Continuous variables with a missing percentage below 40% were subjected to median imputation. No categorical variable with missing values was detected. After imputation, the continuous predictor variables were standardized using z-score.

Algorithms were trained using nested cross-validation with 5 inner folds and 10 outer folds. In the inner fold, hyperparameters of the algorithms were optimized using grid search or random search techniques (in the case of boosting algorithms). Hyperparameters were optimized using the square root of the mean squared error (RMSE).

A total of seven distinct regression algorithms were tested: Lasso Regressor (Tibshirani, 1996), Ridge Regressor (Marquardt & Snee, 1975), Elastic Net (Zou & Hastie, 2005), Extra Tree Regressor (Geurts et al., 2006), LightGBM Regressor (Ke et al., 2017), CatBoost Regressor (Prokhorenkova et al., 2018), and XGBoost Regressor (Chen & Guestrin, 2016). The predictive performance evaluation was based on metrics such as mean squared error (MSE), RMSE, and R^2. Graphical interpretation was performed by comparing the predicted values versus observed values. For variable selection, the Boruta (Kursa & Rudnicki, 2010) technique was applied, which allows for assessing the importance of an original variable by using its shuffled mirrored version. Boruta ranks the features according to three categories where the green area represents variables that have shown significant importance, thus being recommended by the algorithm.

The predictive performance of each variable was assessed by calculating their Shapley Values (Lundberg et al., 2020). We followed the recommendations of the Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD) (Collins et al., 2015). The analysis was conducted in R and Python.

Municipalities Grouping

Municipalities were classified based on the prediction error, which represents the difference between the value expected by the algorithm and the actual observed value of AAPNM. Municipalities with prediction errors falling in the highest decile were categorized as "High," meaning municipalities where the actual value significantly exceeded the expected value. Conversely, municipalities with prediction errors in the lowest decile were classified as "Low", indicating that the observed mortality rate was lower than the expected value. The remaining municipalities were labeled as "Expected."

Furthermore, municipalities that were previously identified as upper outliers in boxplots were designated as "High Outliers," while those identified as lower outliers were classified as "Low Outliers."

Populational filters

The SSP is composed of 645 municipalities, with significant variations regarding population size. Even when adopting the consolidated mortality data from 2010 to 2019, the mortality rates exhibited significant sensitivity to population size, posing challenges to the learning process of the algorithms. In order to overcome this issue, we applied population-based filters based on the classification criteria established by the Brazilian Institute of Geography and Statistics (IBGE) and adopted by the SSP (SSP Government, 2023), which divides municipalities into small-sized I (up to 20,000 inhabitants), small-sized II (20,001-50,000 inhabitants), medium-sized (50,001-100,000 inhabitant), large-sized (100,001-900,000 inhabitants), and metropolis (900,001 inhabitants or more). Based on this classification, we developed specific algorithms without population filter, algorithms for municipalities above 20,000 inhabitants, and algorithm for municipalities above 50,000 inhabitants. The results of the models are presented separately, considering these population-based approaches.

Group Profiling

Based on the algorithm with the best performance, a descriptive analysis of the municipalities from the different categories was conducted. This descriptive analysis took into consideration the predictors selected by the Boruta algorithm. Subsequently, four variables related to primary care were examined to

understand the correlation and statistical differences between the groups classified by the algorithms and health indicators.

The indicators were selected based on primary care data available from the Ministry of Health through the Primary Care Health Information System (SISAB). The focus on primary care was chosen due to its importance for chronic patients (Ministry of Health, 2014). The following indicators were extracted, disaggregated by municipality: Nursing workload in primary care, Medical workload in primary care, Family Health Strategy (ESF) Coverage, Primary Care Coverage.

To analyze the statistical differences between the groups of municipalities classified by the algorithm, Analysis of Variance (ANOVA) tests were performed for each of the four health variables. The null hypothesis assumed statistical equality among the groups, while the alternative hypothesis posited that at least one group differed from the others. For tests where the alternative hypothesis was accepted at a p-value < 0.05, Tukey's test was conducted to identify which groups of municipalities differed statistically from each other.

In order to understand the relationship between observed and expected mortality rates generated by the algorithm, Pearson correlations were calculated between mortality rates and the four health variables.

RESULTS

Exploratory Data Analysis

As shown in Table 1, AAPNM ranged from a minimum of 152.098 per 100,000 inhabitants to a maximum of 606.648, representing approximately four times the minimum value, considering the consolidated period from 2010 to 2019.

Table 1. Descriptive summary of age adjusted premature NCD mortality (AAPNM) and the main predictors in the state of São Paulo.

Variable	Minimum	1st quartile	Median	3rd quartile	Maximum	Standard Deviation
Observed age adjusted premature NCD Mortality/100,000 inhabitants[1]	152.10	307.41	338.99	372.59	606.65	59.38
Predictors[2]						
Average Residence Income in 2010 R$ (White Population)	338.48	637.56	748.95	890.86	2632.90	236.17
Average Residence Income in 2010 R$ (Mixed Population)	244.37	443.84	496.03	557.45	1160.26	96.52
Social benefits for deficient population (General Population)	0.00	32.50	93.00	261.00	105669.00	4216.46
Percent of residences with garbage collected by cleaning service (General Population)	0.01	0.83	0.90	0.95	1.00	0.10
Percent of residences with garbage burned on the property (General Population)	0.00	0.02	0.05	0.10	0.39	0.06
Demographic density (General Population)	3.73	19.67	38.87	110.10	12519.10	1198.31
Percent of urban population (Urban Population Population)	0.25	0.79	0.88	0.95	1.00	0.14
Source of water distribution - sewer or pluvial network (percent of residences for general Population)	0.00	0.00	0.00	0.01	0.22	0.03
Percent of population with twenty-five years and over with incomplete high school (General Population)	0.07	0.13	0.15	0.16	0.26	0.03
Unemployment rate (General Population)	1.33	4.64	6.20	7.69	14.27	2.31
Percent of residence without surrounding afforestation (White Population)	0.00	0.01	0.04	0.17	0.98	0.17
Percent of population in both religious and civil marriage (General Population)	0.24	0.45	0.52	0.57	0.77	0.09
Percent of population in a civil marriage (General Population)	0.07	0.15	0.18	0.21	0.44	0.05
Percent of population in a consensual union (General Population)	0.13	0.25	0.29	0.33	0.49	0.06
Annual Gross domestic product (GDP) per capita (General Population)	4470.44	11057.06	15415.76	22374.39	200186.83	16547.72
Illiteracy Rate (Black Population)	0.00	8.25	12.00	16.40	44.40	6.67

[1] Calculated considering the aggregated deaths from 2010 to 2019.
[2] Considering the final model with feature selection, with basis in 2010

A similar behavior was observed for the main predictors. It is worth noting the racial disparities regarding the variables of average household income. The maximum value for the white population was R$ 2,632.90, contrasting with R$ 1,160.26 for the black population. The analysis of all variables used in the study is available in the Supplementary Appendix A (dataset_statistics.xlsx). ***3.2.***

Outliers Analysis

Considering the whole SSP, we identified 21 municipalities with mortality rates above the upper limit that were then removed from the model. Similarly, eight municipalities had mortality rates below the lower limit. Both outliers above the upper limit and below the lower limit were removed from the model but not excluded from the further analysis.

General Model

The overall model for AAPNM was trained with 616 municipalities using seven different algorithms. After preprocessing, a total of 119 were selected for training the algorithm. The results of the nested cross-validation are presented in Supplementary Appendix A (Table A3, general_nested_results.xlsx). The catboost algorithm had the best predictive performance with an RMSE of 42.171. Figure B2 (Supplementary Appendix B) indicates that the overall model did not demonstrate a good fit, which may be attributed to the high variability of predictors and the outcome among municipalities, as well as the high number of predictors. The use of Boruta resulted in a reduction of 119 to 16 variables. A new training was conducted for Catboost using these 16 variables. The new model yielded an RMSE of 40.842, MSE of 1,668.091, and R^2 of 0.315, indicating improved performance compared to the model with all variables. Figure B1 (Supplementary Appendix B) presents the graph of observed values versus predicted values for the model with variable selection via Boruta.

Populational filter for municipalities with more than 20,000 inhabitants

After applying the populational filter of 20,000 inhabitants, the total number of municipalities decreased from 645 to 244, of which 15 had mortality rates above the upper limit in boxplot analysis. The correlation analysis reduced the number of predictive variables to 105. Table A4 (Supplementary Appendix A, general_nested_results.xlsx) presents the results for the nested cross-validation.

The best algorithm for municipalities above 20,000 inhabitants was the Extra Tree Regressor. After applying Boruta, the number of variables was reduced to 8 (the list is available in the Supplementary Appendix A, data_dictionary.xslx). For the model with 8 variables, the RMSE was 34.822, MSE was 1,212.549, and R^2 was 0.400. Figure B2 of the Supplement B (Supplementary Appendix B) presents the observed values vs predicted values for the complete model (a) and the model with Boruta (b), showing the improvement in model fit with fewer variables.

The best algorithm for municipalities above 20,000 inhabitants was the Extra Tree Regressor. After applying Boruta, the number of variables was reduced to 8 (the list is available in the Supplementary Appendix A, data_dictionary.xslx). For the model with 8 variables, the RMSE was 34.822, MSE was 1,212.549, and R^2 was 0.400. Figure B3 of Supplement B (Supplementary Appendix B) presents the observed values vs predicted values for the complete model (a) and the model with Boruta (b), showing the improvement in model fit with fewer variables.

Populational filter for municipalities with more than 50,000 inhabitants

The best overall predictive models were achieved for the municipalities with more than 50,000 inhabitants. The filter for these municipalities resulted in 124 eligible municipalities, of which 6 had mortality rates above the upper limit. The initial model was trained with 94 variables, and Catboost was the best-performing algorithm, with an RMSE of 39.498 and R^2 of 0.495 (Table 2).

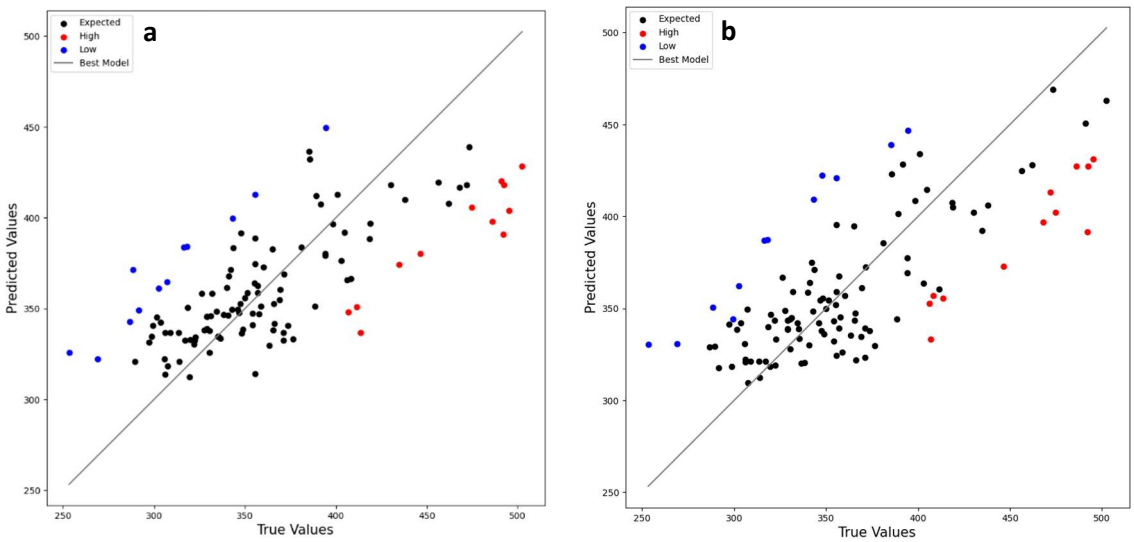
Table 2. Performance of nested cross-validation without Boruta feature selection. Model with populational filter for 50,000 inhabitants and over.

Model	MSE	RMSE	R^2
Catboost Regression	1482.065	38.498	0.495
Extra Tree Regression	1557.074	39.460	0.470
Elastic Net Regression	1663.773	40.789	0.434
LightGBM Regression	1675.885	40.938	0.429
Xgboost Regression	1796.980	42.391	0.388
Lasso Regression	2124.040	46.087	0.277
Ridge Regression	2386.531	48.852	0.187

After applying the Boruta method, the number of variables was reduced to 9, with an RMSE of 37.246, MSE of 1387.228, and R^2 of 0.528. Figure 1 presents the scatter plot of observed values vs. predicted values for the model without Boruta (a) and with Boruta (b). Similar to the overall model and the model for municipalities with 20,000 inhabitants, reducing the number of variables using Boruta resulted in improved predictive performance. For the 50,000 inhabitants and over filter, a total of twelve municipalities were classified as High. Another twelve were classified as 'Low.' Ninety-four fell within the expected range, and six were previously separated as upper outliers, leading to their classification as 'High Outlier.' For the 124 municipalities evaluated, no lower outliers were observed for AAPNM.

The list of variables for the overall model and the model with Boruta is available in the supplementary material (Supplementary Appendix A, data_dictionary.xslx).

Figure 1. Nested cross-validated results for predicted (expected) and observed age adjusted premature NCD mortality (AAPNM) in the municipalities of the state of São Paulo between the years 2010 and 2019. Population-filtered for 50,000 inhabitants and over without (a) and with Boruta feature selection (b).

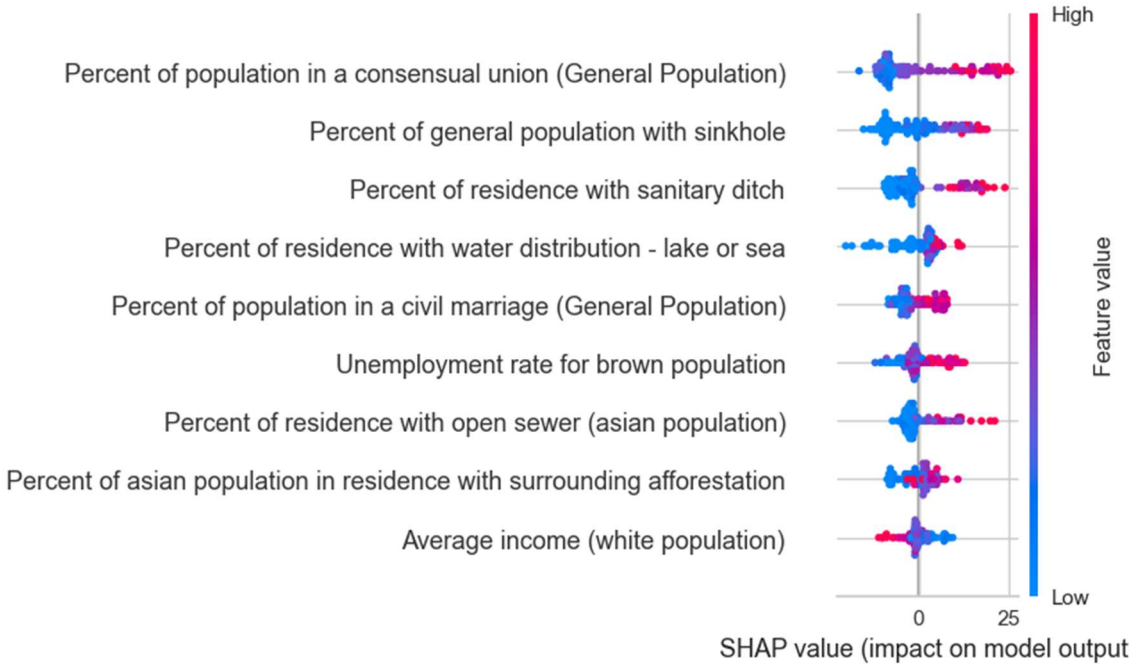


Algorithm's decision path analysis

The Shapley values for the model with 50,000 or more inhabitants (Figure 2) reveal stable union as an important predictor, followed by variables related to garbage collection and water supply. Shapley plots for the general model and

20,000+ model under Boruta feature selection are available in Supplementary Appendix B (Figures B4 and B5, respectively).

Figure 2. Shapley values summary plot for the catboost model filtered by the municipalities with 50,000 inhabitants and over with Boruta feature selection.



Groups Profiling

Based on the municipality categories derived from the machine learning algorithm, we aimed to identify the groups of underachievers (High) and overachievers (Low) concerning the predictors used in the models. According to Table A5 of the Supplement A (table_results.xslx), municipalities classified as High Outliers, those identified by the outlier analysis, had the lowest income levels for the white population and the highest unemployment rates for the mixed population. The observed AAPNM mean was 534.33 deaths per 100,000 inhabitants aged 30 to 69 years, contrasting with 455.17 for the High group, 322.74 for the Low group, and 356.69 for the Expected group. Other infrastructure-related indicators generally showed higher values compared to the other three groups of municipalities. In this regard, the strategy of initially separating municipalities based on outlier analysis proved effective. Although on a smaller scale, municipalities classified as High exhibited similar characteristics to the High Outliers group, particularly low-income levels for the white population and high unemployment rates for the mixed population. The average percentage of households using water from lakes or seas was approximately 5%, higher than

the Expected and Low municipality groups. Other indicators also presented higher values compared to the other groups.

We then assessed the groups of municipalities classified by the algorithms in relation to health variables (Table 3). Significant statistical differences were found only for the variables Nursing workload in primary care and medical workload in primary care, with a p-value < 0.05. For ESF Coverage and Primary Care Coverage, there was no statistical evidence to reject the hypothesis of equality among the High, Low, Expected, and High Outlier groups.

Table 3. Analysis of Variance (ANOVA) results for Health Variables based on algorithm's group classification.

Variable	Eta-squared	F	p-value
Nursing workload in primary care	0.103	3.052	0.031
Medical workload in primary care	0.108	3.055	0.031
ESF Coverage	0.086	1.842	0.143
Primary Care Coverage	0.812	1.401	0.246

Based on the identified differences, a Tukey test was conducted to understand which of the groups showed differences among them (Table A6 of Supplement A, table_results.xlsx). In general, the notable difference was observed only between the Expected and Low groups. Both for Nursing workload in primary care and for Medical workload in primary care, municipalities classified as Expected exhibited higher values for primary care workload compared to municipalities classified as Low.

Figure B6 (Supplementary Appendix B) presents the correlation analysis between health variables and observed and predicted mortalities. A perfect correlation was observed for the variables 'Medical workload in primary care' and 'Nursing workload in primary care,' indicating that, in the evaluated municipalities, medical and nursing care move in the same direction. The variables ESF Coverage and Primary Care Coverage also showed a strong correlation (0.88), which was expected as they are aligned programs aimed at primary healthcare. When assessing the correlation between health variables and mortalities, a strong correlation was not observed. However, this behavior is mainly driven by municipalities classified as 'Expected.'

Figure B7 (Supplementary Appendix B) presents the correlation matrix for municipalities classified as 'High' by the algorithm. It is noticeable that ESF coverage and Primary Care coverage exhibit correlations of 0.53 and 0.48 with predicted mortality, indicating that in municipalities with higher mortality rates, there is greater coverage of primary care. Additionally, a negative correlation is highlighted between medical and nursing workloads and primary care and ESF coverages. For municipalities classified as Low (Supplementary Appendix B, Figure B8), the scenario reverses, with a negative correlation between mortalities and primary care coverages.

DISCUSSION

The study identified municipalities in the SSP with mortality rates for NCD above or below the expected levels, considering their demographic, socioeconomic, and infrastructure characteristics. The use of machine learning, associated with profiling statistical resources, can become an important tool for NCD management and surveillance, especially in public health.

We found that using municipalities with larger populations leads to higher model stability, which may be associated with the sensitivity of mortality indicators in smaller areas. Previous studies have adopted similar modeling approaches by applying populational filters, nested cross-validation, and sociodemographic variables to predict expected values for life expectancy in Brazil (Chiavegatto-Filho et al., 2018). Zhang et al (2019) collected colorectal cancer incidence data from different sources and used the average annual change rate to analyze temporal trends in China. The authors developed regression models to assess tendencies in colorectal incidence and mortality, predicting continued increase in colorectal cancer cases and deaths in China until 2025. Ryzhov et al. (2020) applied logistic models to predict cancer incidence in Ukraine by 2022. May et al. (2019) developed the Intermountain Chronic Disease Model (ICHRON), a model based on laboratory parameters to predict the onset of chronic diseases in primary care patients. This model, named ICHRON, showed good discrimination and long-term risk prediction capability for cardiovascular and cardiopulmonary diseases. The researchers emphasized the importance of this model as a clinical

decision support tool in primary care, providing information for the control and management of chronic diseases.

The variables used as predictors for our study show that the performance of municipalities in the management of NCD tends to be interconnected with social, economic, and infrastructure factors, aligning with the hypotheses of social determinants of health (Cockerham et al., 2017). Based on the proposed model, municipalities with low income, high unemployment rates, and unconventional sources of water and waste disposal appear to be consistent with higher-than-expected mortality rates. This underscores the significance of addressing these socio-economic and infrastructural disparities in efforts to improve NCD outcomes and overall public health.

Currently, the main strategies in the literature for predicting the incidence, prevalence, or mortality of NCD involve trend analysis or future prediction based on conventional and classical statistical models. Our study helps to fill a literature gap by using ML algorithms as a tool for supporting the management of NCD at a surveillance and health policy level. However, some limitations need to be highlighted. First, we used data from the 2010 population data, which is the last Brazilian census with fully available data, but its results could be outdated. Second, there may be consistent underreporting of premature NCD deaths on official data (Malta et al., 2019). Lastly, we found better predictive performance for models with fewer municipalities, which may limit the scope of the study if higher performance is preferred instead of broader coverage.

In conclusion, our study has highlighted a few challenges regarding the complex interaction between social factors, healthcare access, and NCD mortality in municipalities. These results can provide valuable insights into potential determinants of NCD mortality and underscore the need for further investigation. Future research endeavors should focus on empirically testing these results to gain a deeper understanding of the underlying mechanisms driving NCD mortality variations across municipalities. Such investigations hold the potential to inform more targeted public health strategies and interventions aimed at reducing NCD burden and improving healthcare outcomes in diverse community settings.

References

Brasil. Ministério da Saúde. (2023). Mortalidade São Paulo. TabNet: Indicadores de saúde [Mortality São Paulo. TabNet: Health indicators]. Brasília: Ministério da Saúde, Brasil. Retrieved from http://tabnet.datasus.gov.br/cgi/deftohtm.exe?sim/cnv/obt10sp.def

Brasil. Ministério da Saúde. (2014). Estratégias para o cuidado da pessoa com doença crônica. Cadernos de Atenção Básica 35. Retrieved from: https://bvsms.saude.gov.br/bvs/publicacoes/estrategias_cuidado_pessoa_doenca_cronica_cab35.pdf

Brasil. Ministério do Desenvolvimento Social e Combate à Fome (2010). Benfícios Sociais - São Paulo. TabNet: Indicadores de saúde [Mortality São Paulo. TabNet: Health indicators]. Brasília, Brasil. Retrieved from http://www.ipeadata.gov.br/Default.aspx

Budreviciute A, Damiati S, Sabir DK, Onder K, Schuller-Goetzburg P, Plakys G, Katileviciute A, Khoja S, Kodzius R. Management and Prevention Strategies for Non-communicable Diseases (NCD) and Their Risk Factors. Front Public Health. 2020 Nov 26;8:574111. doi: 10.3389/fpubh.2020.574111. PMID: 33324597; PMCID: PMC7726193.

Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).

Chiavegatto-Filho, A. D. P., Dos Santos, H. G., do Nascimento, C. F., Massa, K., & Kawachi, I. (2018). Overachieving municipalities in public health: a machine-learning approach. Epidemiology, 29(6), 836-840.

Cockerham, W. C., Hamby, B. W., & Oates, G. R. (2017). The social determinants of chronic disease. American journal of preventive medicine, 52(1), S5-S12.

Collins, G. S., Reitsma, J. B., Altman, D. G., & Moons, K. G. (2015). Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) the TRIPOD statement. Circulation, 131(2), 211-219.

Dapper, S. N., Spohr, C., & Zanini, R. R. (2016). Poluição do ar como fator de risco para a saúde: uma revisão sistemática no estado de São Paulo. Estudos Avançados, 30, 83-97.

Davenport, T., & Kalakota, R. (2019). The potential for artificial intelligence in healthcare. Future healthcare journal, 6(2), 94.

Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. Machine learning, 63, 3-42.

Instituto Brasileiro de Geografia e Estatística (IBGE). (2010). Censo Demográfico 2010 [2010 Census]. Retrieved from https://www.ibge.gov.br/censo2010/

Instituto Brasileiro de Geografia e Estatística (IBGE). (2023). Panorama Estados. São Paulo. Retrieved from https://cidades.ibge.gov.br/brasil/sp/panorama

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T. Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems, 30.

Kursa, M. B., & Rudnicki, W. R. (2010). Feature selection with the Boruta package. Journal of statistical software, 36, 1-13.

Lones, M. A. (2021). How to avoid machine learning pitfalls: a guide for academic researchers. arXiv preprint arXiv:2108.02497.

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., ... & Lee, S. I. (2020). From local explanations to global understanding with explainable AI for trees. Nature machine intelligence, 2(1), 56-67.

Malta, D. C., Andrade, S. S. C. D. A., Oliveira, T. P., Moura, L. D., Prado, R. R. D., & Souza, M. D. F. M. D. (2019). Probability of premature death for chronic non-communicable diseases, Brazil and Regions, projections to 2025. Revista Brasileira de Epidemiologia, 22.

Marquardt, D. W., & Snee, R. D. (1975). Ridge regression in practice. The American Statistician, 29(1), 3-20.

May, H. T., Lappé, D. L., Knowlton, K. U., Muhlestein, J. B., Anderson, J. L., & Horne, B. D. (2019, July). Prediction of long-term incidence of chronic

cardiovascular and cardiopulmonary diseases in primary care patients for population health monitoring: the Intermountain Chronic Disease Model (ICHRON). In Mayo Clinic Proceedings (Vol. 94, No. 7, pp. 1221-1230). Elsevier.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: unbiased boosting with categorical features. Advances in neural information processing systems, 31.

Ryzhov, A., Bray, F., Ferlay, J., Fedorenko, Z., Goulak, L., Gorokh, Y., ... & Znaor, A. (2020). Recent cancer incidence trends in Ukraine and short-term predictions to 2022. Cancer Epidemiology, 65, 101663.

São Paulo State Government. (2023). Informações Socioterritoriais [Socioterritorial Information]. Retrieved from https://www.desenvolvimentosocial.sp.gov.br/vigilancia-socioassistencial/informacoes-socioterritoriais/

São Paulo State Health Department (SES/SP). (2023). Matriz de indicadores. Retrieved from http://www.saude.sp.gov.br/links/matriz

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological), 58(1), 267–288.

Travassos, C., de Oliveira, E. X., & Viacava, F. (2006). Desigualdades geográficas e sociais no acesso aos serviços de saúde no Brasil: 1998 e 2003. Ciência & Saúde Coletiva, 11, 975-986.

World Health Organization. (2010). International Statistical Classification of Diseases and Related Health Problems (10th Revision), Volume 2: Instruction Manual. Retrieved from https://icd.who.int/browse10/Content/statichtml/ICD10Volume2_en_2010.pdf

World Health Organization. (2022). Noncommunicable diseases: Key facts. Retrieved July 13, 2023, from https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases

World Health Organization. (2023). Premature mortality from noncommunicable disease. Retrieved July 13, 2023, from https://www.who.int/data/gho/indicator-metadata-registry/imr-details/3411

Zerbini, T., Ridolfi, A. D. A. C., da Silva, A. C. C. G., & Rocha, L. E. (2009). Trânsito como fator estressor para os trabalhadores. Saúde Ética & Justiça, 14(2), 77-83.

Zhang, L., Cao, F., Zhang, G., Shi, L., Chen, S., Zhang, Z., … & Ma, T. (2019). Trends in and predictions of colorectal cancer incidence and mortality in China from 1990 to 2025. Frontiers in Oncology, 9, 98.

Zou, H., & Hastie, T. (2005). Regularization and Variable Selection via the Elastic Net. Journal of the Royal Statistical Society. Series B (Statistical Methodology), 67(2), 301–320. http://www.jstor.org/stable/3647580

Considerações Finais e Conclusão

O presente estudo analisou a aplicação de algoritmos de ML em problemas relacionados às DCNT, com o objetivo de desenvolver ferramentas que orientem o cuidado, a promoção e a vigilância em saúde. Diante da urgente necessidade por frear o avanço das DCNT, especialmente nas populações entre 30 e 69 anos, foram aplicadas diferentes técnicas de aprendizado, com o objetivo de analisar algumas das potenciais contribuições da inteligência artificial na saúde pública.

No primeiro artigo, a revisão sistemática de literatura, embora focada em predição de hipertensão arterial, demonstrou o potencial da utilização de algoritmos de ML em desfechos crônicos. Ao mesmo tempo, identificou-se uma necessidade de amadurecimento científico da área, com a incorporação de boas práticas preditivas, criteriosidade na seleção de preditores, definição clara do desfecho predito, além da reflexão sobre a relevância e aplicabilidade dos algoritmos desenvolvidos. O desenvolvimento desse artigo foi de grande importância para nortear as boas práticas adotadas na construção dos algoritmos de ML, supervisionados ou não supervisionados, orientando a construção dos artigos seguintes.

O segundo artigo analisou a aplicação de algoritmos de ML de aprendizado supervisionado na predição do risco de óbito por câncer entre 12 e 24 meses após o diagnóstico, em pacientes do Estado de São Paulo. Os resultados obtidos apontam o potencial em se utilizar ferramentas prognósticas no cuidado de pacientes com câncer. Os algoritmos desenvolvidos apresentaram boa performance preditiva (AUC-ROC de 0,946), indicando a viabilidade em se incorporar essas ferramentas na prática clínica. Variáveis como o estadiamento clínico, a morfologia do câncer e a faixa etária demonstraram-se como importantes preditores. No entanto, embora o estudo tenha apresentado boas métricas de desempenho, destaca-se que a incorporação de algoritmos de ML na saúde deve ser feita de maneira cuidadosa. Ademais, apesar de o RHC-FOSP possuir pacientes de todo o Estado de São Paulo, para que os resultados possam ser extrapolados para toda a população, é necessária a realização de uma avaliação do perfil dos participantes presentes

na base de dados, evitando-se vieses e sobreajustes para a população específica.

O terceiro artigo buscou fornecer uma análise da distribuição regional dos municípios do estado de São Paulo baseando-se no perfil de morbimortalidade por DCNT entre os anos de 2010 e 2019. Para isso, foram utilizados algoritmos de ML não supervisionado, especificamente os de cluster, para identificar grupos de municípios com perfis epidemiologicamente semelhantes entre si. Embora o Estado de São Paulo seja, desde o ano de 2012, dividido em 17 regiões de saúde, o estudo demonstrou que os critérios administrativos e epidemiológicos adotados à época podem não refletir o cenário observado no decorrer da década. No entanto, reforça-se que esse trabalho não buscou refutar o desenho atual das redes regionais de atenção à saúde, mas sim desenvolver um método que forneça insumos para uma nova atualização da distribuição regional do estado, que está prevista para breve segundo a Secretaria de Estado da Saúde de São Paulo.

No quarto e último artigo, foram desenvolvidos algoritmos supervisionados para valores contínuos, com o objetivo de construir um método para auxiliar na avaliação da performance da gestão das doenças crônicas não transmissíveis dos municípios do Estado de São Paulo. Os resultados demonstraram que, embora seja possível utilizar algoritmos como ferramenta de auxílio na avaliação da gestão de saúde, existe um importante trade-off entre performance e abrangência dos modelos: os algoritmos que contemplaram todos os municípios do Estado não tiveram bom desempenho, com RMSE de 40,842, MSE de 1668,091 e R^2 de 0,315. Após a aplicação do filtro de base populacional para municípios acima de 20.000 habitantes, o RMSE, MSE e R^2 aumentaram para 34,822, 1212,549 e 0,400, respectivamente. Para os municípios com mais de 50.000 habitantes, foram observados um RMSE de 37,246, MSE de 1387,228, R^2 de 0.528. Nesse sentido, uma das limitações a ser considerada é a possibilidade de viés nos resultados devido à seleção da base populacional, uma vez que o filtro adotado pode levar à exclusão de informações importantes provenientes de municípios com populações menores. Além disso, reforça-se que a implementação desses algoritmos requer uma interpretação cuidadosa dos resultados, considerando a complexidade das dinâmicas de saúde locais, a

influência de fatores socioeconômicos e culturais, bem como a evolução das estratégias de gestão ao longo do tempo. Embora os algoritmos possam fornecer resultados relevantes, eles devem ser considerados como parte de um conjunto de ferramentas para a aprimoração da gestão de DCNT, integrando conhecimentos especializados e considerações contextuais para a tomada de decisão.

Espera-se que esta tese contribua com o desenvolvimento científico no campo da saúde pública, ao explorar as possibilidades da IA na abordagem de desafios complexos. Ao analisar diferentes problemas relacionados às DCNT, este estudo buscou destacar novas formas de aprimorar a gestão, promoção e vigilância da saúde. Acredita-se que, ao enfrentar as complexidades e limitações apontadas, as descobertas desta pesquisa possam enriquecer a compreensão das dinâmicas de saúde, estimulando discussões interdisciplinares e contribuindo para a tomada de decisões embasadas em evidências empíricas.

Este trabalho apresenta contribuições que têm o potencial de transformar o futuro da vigilância de DCNT. Algumas maneiras pelas quais esse trabalho pode influenciar positivamente a vigilância em saúde de DCNT incluem:

- Integração de tecnologias avançadas: a aplicação bem-sucedida de algoritmos de ML destaca a viabilidade e a eficácia de integrar tecnologias avançadas na vigilância em saúde. Esse exemplo prático pode incentivar a adoção gradual de abordagens inovadoras para análise de dados e predição de desfechos relacionados às DCNT, contribuindo para uma gestão de saúde que conte cada vez mais com a algoritmos para orientar a tomada de decisão e o desenho de políticas de saúde.
- Orientação para tomada de decisão clínica: os modelos desenvolvidos, particularmente no segundo artigo, têm o potencial de orientar a tomada de decisão clínica, fornecendo insights prognósticos valiosos e apoiando profissionais de saúde na gestão eficaz de pacientes com DCNT, no caso, o câncer. Isso pode resultar em intervenções mais rápidas e adaptadas às necessidades individuais dos pacientes, com o potencial de

melhorar os desfechos e, consequentemente, a situação de saúde do estado de São Paulo.

- Revisão e atualização da distribuição regional: o terceiro artigo, ao analisar a distribuição regional dos municípios com base no perfil de morbimortalidade por DCNT, oferece insumos para uma potencial revisão e atualização da distribuição regional de saúde do estado de São Paulo. Essa análise mais granular pode levar a estratégias de vigilância mais focadas e ajustadas às realidades locais.
- Conscientização sobre limitações e desafios: ao destacar cuidadosamente as limitações e desafios associados à implementação de algoritmos em larga escala, os quatro artigos contribuem para uma conscientização crítica sobre a necessidade de avaliações criteriosas e considerações cuidadosas na interpretação dos resultados, buscando promover uma implementação mais informada e ética das ferramentas de ML na vigilância em saúde.
- Estímulo a pesquisas futuras e colaborações Interdisciplinares: ao demonstrar a aplicabilidade e os benefícios da IA na vigilância em DCNT, este trabalho pode estimular pesquisas futuras e promover colaborações interdisciplinares entre a academia, profissionais de saúde e formuladores de políticas públicas. Essa colaboração pode acelerar a inovação e a implementação prática de soluções baseadas em dados para enfrentar os desafios das DCNT.

Ao destacar a necessidade contínua de abordagens sensíveis e avaliações criteriosas, esta tese não apenas busca promover a aplicação ética e eficaz da IA na saúde, mas também pretende fomentar um ambiente de pesquisa no qual a inovação tecnológica é aliada ao rigor técnico. A expectativa é que esta tese não apenas estimule futuras pesquisas no campo, mas também impulsione a colaboração entre a academia, profissionais de saúde e formuladores de políticas públicas, na área de IA e ML. Assim, almeja-se que os resultados aqui

apresentados contribuam para o conhecimento científico e se traduzam em medidas práticas que beneficiem a saúde da população.

Referências

1. Cesse EAP. Epidemiologia e determinantes sociais das doenças crônicas não transmissíveis no Brasil (Doctoral dissertation); 2007. Disponível em: https://www.arca.fiocruz.br/bitstream/handle/icict/3905/000006.pdf?sequence=2&isAllowed=y
2. World Health Organization. Noncommunicable diseases. Fact Sheets; 2022. Disponível em: https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases
3. Pan American Health Organization. The burden of noncommunicable diseases in the Region of the Americas, 2000-2019. ENLACE data portal; 2021. Disponível em: https://www.paho.org/en/enlace/burden-noncommunicable-diseases
4. Ministério da Saúde (BR). Sistema de informação sobre mortalidade (SIM); 2023. Disponível em: http://tabnet.datasus.gov.br/cgi/tabcgi.exe?sim/cnv/obt10br.def
5. Ministério da Saúde (BR). Plano de Ações Estratégicas para o Enfrentamento das Doenças Crônicas e Agravos Não Transmissíveis no Brasil; 2021. Disponível em: https://www.gov.br/saude/pt-br/centrais-de-conteudo/publicacoes/svsa/doencas-cronicas-nao-transmissiveis-dcnt/09-plano-de-dant-2022_2030.pdf/view
6. Singh A, Thakur N, Sharma A. A review of supervised machine learning algorithms. In 2016 3rd international conference on computing for sustainable global development (INDIACom). IEEE access; 2016. 1310-1315
7. Javaid M, Haleem A, Singh RP, Suman R, Rab, S. Significance of machine learning in healthcare: Features, pillars and applications. International Journal of Intelligent Networks; 2022; 3:58-73. https://doi.org/10.1016/j.ijin.2022.05.002.
8. Alanazi A. Using machine learning for healthcare challenges and opportunities. Informatics in Medicine Unlocked; 2022; 30, 100924. https://doi.org/10.1016/j.imu.2022.100924.

9. Jiang T, Gradus JL, Rosellini AJ. Supervised machine learning: a brief primer. *Behavior Therapy*; 2022; *51*(5):675-687. https://doi.org/10.1016/j.beth.2020.05.002.
10. Bzdok D, Krzywinski M, Altman N. Machine learning: supervised methods. Nature methods; 2018; 15(1), 5. https://doi.org/10.1038/nmeth.4551.
11. Alloghani M, Al-Jumeily D, Mustafina J, Hussain A, Aljaaf AJ. A systematic review on supervised and unsupervised machine learning algorithms for data science. Supervised and unsupervised learning for data science; 2020; 3-21. https://doi.org/10.1007/978-3-030-22475-2_1.
12. Usama M, Qadir J, Raza A, Arif H, Yau K A, Elkhatib Y, Hussain A, Al-Fuqaha A. Unsupervised machine learning for networking: Techniques, applications and research challenges. IEEE access; 2019; 7: 65579-65615. https://doi.org/10.1109/ACCESS.2019.2916648.
13. World Health Organization. *Preventing noncommunicable diseases (NCDs) by reducing environmental risk factors* (No. WHO/FWC/EPE/17.01). World Health Organization; 2017.
14. Silva GF, Fagundes TP, Teixeira BC, Chiavegatto-Filho, AD. Machine learning for hypertension prediction: a systematic review. Current Hypertension Reports; 2022; 24(11):523-533. https://doi.org/10.1007/s11906-022-01212-6.
15. Moons KG, Altman DG, Reitsma JB, Ioannidis JP, Macaskill P, Steyerberg EW, Vickers AJ, Ransohoff DF, Collins GS. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. Ann Intern Med. 2015 Jan 6;162(1):W1-73. https://doi.org/10.7326/M14-0698. PMID: 25560730.
16. van de Schoot R, de Bruin J, Schram R. et al. An open source machine learning framework for efficient and transparent systematic reviews. Nat Mach Intell; 2021; 3:125–133. https://doi.org/10.1038/s42256-020-00287-7
17. Silva GFS, Duarte LS, Shirassu MM, Peres SV, de Moraes MA, Chiavegatto-Filho, A. Machine learning for longitudinal mortality risk prediction in patients with malignant neoplasm in são paulo, Brazil. Artificial Intelligence in the Life Sciences; 2022; 3. https://doi.org/10.1016/j.ailsci.2023.100061.