LARISSA OLIVEIRA CHAVES

PERFIL SOCIODEMOGRÁFICO, CLÍNICO, DE ESTILO DE VIDA E DE CONSUMO ALIMENTAR DE UMA POPULAÇÃO CARDIOPATA (ESTUDO DICA BR): UMA INVESTIGAÇÃO BASEADA EM *MACHINE LEARNING*

Tese apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Ciência da Nutrição, para obtenção do título de *Doctor Scientiae*.

Orientador: Josefina Bressan

Coorientadores: Rodrigo Siqueira Batista
Ana Luiza Gomes Domingos

VIÇOSA – MINAS GERAIS
2021

Ficha catalográfica elaborada pela Biblioteca Central da Universidade Federal de Viçosa - Campus Viçosa

T
O48p
2021

Chaves, Larissa Oliveira, 1986-
Perfil sociodemográfico, clínico, de estilo de vida e de consumo alimentar de uma população cardiopata (Estudo DICA Br): uma investigação baseada em machine learning / Larissa Oliveira Chaves. – Viçosa, MG, 2021.
1 tese eletrônica (159 f.): il. (algumas color.).

Inclui anexos.
Orientador: Josefina Bressan.
Tese (doutorado) - Universidade Federal de Viçosa.
Referências bibliográficas: f. 136-146.
DOI: https://doi.org/10.47328/ufvbbt.2021.080
Modo de acesso: World Wide Web.

1. Doenças cardiovasculares. 2. Ingestão de Alimentos. 3. Aprendizado do computador. 4. Micronutrientes. I. Universidade Federal de Viçosa. Departamento de Nutrição e Saúde. Programa de Pós-Graduação em Ciência da Nutrição. II. Título.

CDD 22. ed. 616.12

Bibliotecário(a) responsável: Renata de Fátima Alves CRB6/2578

LARISSA OLIVEIRA CHAVES

PERFIL SOCIODEMOGRÁFICO, CLÍNICO, DE ESTILO DE VIDA E DE CONSUMO ALIMENTAR DE UMA POPULAÇÃO CARDIOPATA (ESTUDO DICA BR): UMA INVESTIGAÇÃO BASEADA EM *MACHINE LEARNING*

Tese apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Ciência da Nutrição, para obtenção do título de *Doctor Scientiae*.

APROVADA: 17 de setembro de 2021.

Assentimento:

Larissa O. Chaves

Larissa Oliveira Chaves
Autora

Josefina Bressan

Josefina Bressan
Orientadora

Dedico este trabalho a Jesus Cristo, que me deu força e coragem para continuar.

Aquele que colocou anjos na minha vida para que eu pudesse chegar até aqui!

Aquele que está sempre ao meu lado derramando seu Espírito Santo sobre mim!

AGRADECIMENTOS

Primeiramente a Deus, pela minha vida e por sempre me guiar, iluminando todos os meus passos e estando sempre ao meu lado. Obrigado meu Senhor! A Nossa Senhora Aparecida pela intercessão e proteção em todos os momentos.

Aos meus pais, Wilson e Clarinda, pelo amor incondicional, exemplos de força e dedicação. Com todas as dificuldades nunca mediram esforços para que eu e meus irmãos estudássemos e tivéssemos uma formação de qualidade, mostrando a importância dos estudos em nossas vidas. Essa vitória é toda de vocês, e espero de coração poder retribuir pelo menos um pouquinho de tudo o que já fizeram pela gente. Amo vocês!!!

Aos meus irmãos, Thiago e Jaísa, por todo o carinho, irmandade e a fiel parceria de sempre, com vocês com certeza a caminhada foi mais leve.

As minhas lindas avós, exemplos de amor, amo muito vocês.

Ao meu esposo, Rafael, por todo seu amor, carinho, cuidado e dedicação. Um companheiro exemplar, sempre me apoiando, e nos momentos de cansaço e desanimo sempre com palavras de força e apoio. Obrigado meu amor, por ser esse homem humano fantástico que você é.

A minha orientadora, Josefina Bressan, pela confiança em mim depositada, pelas oportunidades desde a graduação e mestrado, anos de muito trabalho e aprendizado. Agradeço por todo seu apoio, ensinamentos, carinho e por me ajudar a concretizar esse sonho. Tenho muita admiração e respeito pelo seu trabalho e pela pessoa que é! Obrigada por tudo!

Aos meus coorientadores e amigos, Ana Luiza e Daniel, a quem dedico em especial esse trabalho, por toda orientação e dedicação, com certeza parte do que me tornei na vida acadêmica devo a vocês. Sempre prontos para ajudar, inúmeras foram às vezes em que me socorreram, e quantos foram os ensinamentos. Obrigado por toda essa caminhada em que estivemos juntos, pela leveza e amizade, só tenho a agradecer. Nossa equipe sempre estará em meu coração. Saiba que tenho grande admiração por vocês.

Ao meu coorientador, Rodrigo Siqueira Batista, por ter encabeçado este projeto, por ter compartilhado seu conhecimento e nos dado força e coragem para fazê-lo. Também agradeço por todo o apoio, carinho, atenção, contribuições e

disponibilidade em ajudar, principalmente por ter me acolhido tão bem para que pudéssemos realizar este trabalho. Obrigado!

Ao Professor Fábio Cerqueira do Departamento de Ciência da Computação da UFV e do Departamento de Engenharia da Produção da Universidade Federal Fluminense por toda contribuição nas análises e na escrita dos artigos científicos e pela parceria com o nosso projeto. Muito obrigado!

A toda equipe DICA Br que desenvolveram o projeto DICA e coletaram os dados que hoje fazem parte da minha tese. E pela parceria e colaboração nos artigos científicos, em especial a Bernardete Weber, Ângela Cristine Bersh-Ferreira, Camila Ragne Torreglosa e Aline Marcadenti. Muito obrigado!

Aos amigos do LAMECC, pela amizade, ajuda e horas de trabalhos compartilhadas. Não posso esquecer os inúmeros cafés da tarde que proporcionavam momentos de descontração e altas risadas deixando os nossos dias mais leves e alegres. O meu muito obrigado!

A amiga Natália Liberto, nossa querida técnica do LAMECC, eu agradeço toda a amizade, palavras de conforto, apoio e preocupação. Nunca vou me esquecer de todos os momentos em que você se fez presente, sou eternamente grata.

Aos voluntários do estudo DICA Br pela valiosa colaboração e comprometimento. Sem a ajuda de vocês a realização deste sonho não seria possível!

Ao Hospital do Coração pela coordenação do DICA-Br e ao Ministério da Saúde que por intermédio do Programa de Apoio ao Desenvolvimento Institucional do Sistema Único de Saúde (PROADI-SUS) financiou o estudo.

Ao Departamento de Nutrição e Saúde e ao Programa de Pós-Graduação em Ciência da Nutrição, pelo apoio e todos os ensinamentos ao longo desses anos.

A Universidade Federal de Viçosa, que me acolheu de braços abertos e proporcionou momentos inesquecíveis, desde a graduação, mestrado, até os dias de hoje. Em especial ao Departamento de Nutrição e Saúde, todos os professores e funcionários que contribuíram muito para a minha formação acadêmica e pessoal.

A CAPES pelo primeiro ano de concessão da bolsa de doutorado. Ao CNPq e FAPEMIG por toda ajuda de sempre.

Ao Instituto Federal do Mato Grosso (IFMT) que mesmo após tomar posse do meu concurso me permitiu ir a UFV por muitas vezes para realizar as atividades do doutorado, e por ter me liberado para prestar colaboração técnica no Instituto

Federal Sudeste de Minas Gerais (IFSUDESTEMG), a quem agradeço também por me liberar para assistir a disciplina de seminários, o que permitiu que eu pudesse dar andamento e concluir meu doutorado. Fica aqui toda minha gratidão as duas Instituições.

A todos meus queridos amigos pelo carinho, apoio, torcida, consolo, conselhos e risadas. Muito obrigado pela grande amizade e por fazerem parte da minha vida!

A toda minha família, pelo amor, amizade e apoio em todos os momentos.

Por fim, obrigado a todos que direta ou indiretamente contribuíram para transformar este sonho em realidade. Vocês são muito especiais!

“É preciso que você se torne a mudança que deseja ver no mundo”.

(Mahatma Gandhi)

RESUMO

CHAVES, Larissa Oliveira, D.Sc., Universidade Federal de Viçosa, setembro de 2021. **Perfil sociodemográfico, clínico, de estilo de vida e de consumo alimentar de uma população cardiopata (Estudo DICA Br): uma investigação baseada em *machine learning*.** Orientador: Josefina Bressan. Coorientadores: Rodrigo Siqueira Batista e Ana Luiza Gomes Domingos.

As doenças cardiovasculares (DCV) representam a principal causa de morte no mundo, tendo um impacto financeiro na qualidade de vida dos indivíduos. Para a abordagem deste importante problema de saúde pública, técnicas de *Machine Learning* (ML) na área da saúde têm sido utilizadas para a prevenção e previsão de doenças. O objetivo desta tese foi investigar uma população constituída por cardiopatas em atenção secundária, com foco nas suas características sociodemográficas, clínicas, de estilo de vida e de consumo alimentar, utilizando algoritmos de ML e estatísticas clássicas. Este é um estudo transversal com dados basais do "*Brazilian Cardioprotective Nutritional Program - BALANCE Program*" que incluiu 1990 pacientes. Os seguintes dados foram obtidos por meio de questionários: características socioeconômicas, clínicas e comportamentais, além de avaliação bioquímica dos principais marcadores de risco cardiovascular. Para a análise de agrupamento foi avaliado os algorithms *k-means, hierarchical agglomerative clustering, expectation-maximization,* e *spectral clustering.* A revisão sistemática foi conduzida de acordo com o PRISMA e registrado no Banco de dados PROSPERO. Uma busca na literatura foi realizada usando cinco bancos de dados eletrônicos e 36 estudos originais foram incluídos na revisão. Dos 1990 participantes 58,4% eram do sexo masculino, com mediana de idade de 63 anos. Além disso, 53,8% eram ex-fumantes e 65% não praticavam atividade física. Foram encontrados quatro grupos, dois compostos exclusivamente por homens e dois por mulheres. Foi observada nos grupos masculinos uma maior renda; escolaridade; fumantes e ex-fumantes; praticantes de atividade física; maior consumo de calorias, ácidos graxos poliinsaturados e ômega 6 e um menor consumo de ácidos graxos saturados. Além disso, observou-se presença mais frequente de doença arterial coronariana e infarto agudo do miocárdio. Os grupos compostos por mulheres apresentaram mais obesidade, diabetes *mellitus*, hipertensão arterial sistêmica, dislipidemia e mais

fatores de risco para DCV. Indivíduos com dois ou mais eventos cardiovasculares, comparados aos indivíduos com apenas um evento, apresentavam menor renda, hábito de fumar, concentrações elevadas de glicemia e baixa lipoproteína de alta densidade, maior razão cintura/estatura, maior consumo de alimentos culinários processados e menor consumo de fibras. Além disso, observou-se também baixa ingestão de cobre, magnésio, selênio, vitaminas B12 e C. Em conclusão, os resultados revelaram diferenças relacionadas ao sexo e ao uso de hipoglicemiantes nas DCV que podem levar a diversos eventos cardiovasculares. Além disso, as técnicas de ML podem ser uma ferramenta robusta para explorar padrões e relações aplicadas ao problema de DCV. Ademais, os resultados sugerem relações importantes entre as baixas concentrações de micronutrientes e um perfil metabólico e alimentar mais alterado nos indivíduos com mais de um evento cardiovascular. O uso de ML na nutrição é recente e desafiador, portanto, é encorajado que mais estudos sejam realizados relacionando esses temas para o desenvolvimento de programas de reeducação alimentar e políticas públicas.

Palavras-chave: Doença cardiovascular. Consumo alimentar. Micronutrientes. Aprendizado de máquina. Algoritmos de agrupamento. Algoritmos supervisionados e não supervisionados.

ABSTRACT

CHAVES, Larissa Oliveira, D.Sc., Universidade Federal de Viçosa, September 2021. **Sociodemographic, clinical, lifestyle and food consumption profiles of a population with heart disease (DICA Br Study): an investigation based on *machine learning***. Adviser: Josefina Bressan. Co-advisers: Rodrigo Siqueira Batista and Ana Luiza Gomes Domingos.

Cardiovascular diseases (CVD) represent the leading cause of death in the world, having a financial impact on the quality of life of individuals. To address this important public health problem, Machine Learning (ML) techniques in healthcare have been used for disease prevention and prediction. The aim of this thesis was to investigate a population consisting of heart patients in secondary care, focusing on their sociodemographic, clinical, lifestyle and food consumption characteristics, using ML algorithms and classical statistics. This is a cross-sectional study with baseline data from the "Brazilian Cardioprotective Nutritional Program - BALANCE Program" that included 1990 patients. The following data were obtained through questionnaires: socioeconomic, clinical and behavioral characteristics, and biochemical evaluation of the main cardiovascular risk markers. For cluster analysis, the *k-means, hierarchical agglomerative clustering, expectation-maximization,* and *spectral clustering algorithms* were used. The systematic review was conducted according to PRISMA and registered in the PROSPERO database. A literature search was conducted using five electronic databases and 36 original studies were included in the review. Of the 1990 participants 58.4% were male, with a median age of 63 years. In addition, 53.8% were former smokers and 65% did not practice physical activity. Four groups were found, two composed exclusively of men and two of women. It was observed in the male groups a higher income; education; smokers and ex-smokers; physical activity practitioners; higher consumption of calories, polyunsaturated fatty acids, and omega 6, and a lower consumption of saturated fatty acids. In addition, the presence of coronary artery disease and acute myocardial infarction was more frequent. The groups composed of women had more obesity, diabetes mellitus, hypertension, dyslipidemia, and more risk factors for CVD. Individuals with two or more cardiovascular events, compared to individuals with only one event, had lower income, smoking habits, high blood glucose and low high-density lipoprotein

concentrations, higher waist-to-height ratios, higher consumption of processed culinary foods, and lower fiber intake. In addition, low intakes of copper, magnesium, selenium, vitamins B12 and C were also observed. In conclusion, the results revealed differences related to gender and the use of hypoglycemic agents in CVD that can lead to various cardiovascular events. Moreover, ML techniques can be a robust tool to explore patterns and relationships applied to the CVD problem. Furthermore, the results suggest important relationships between low micronutrient concentrations and a more altered metabolic and dietary profile in individuals with more than one cardiovascular event. The use of ML in nutrition is recent and challenging, so it is encouraged that further studies are conducted relating these issues for the development of dietary re-education programs and public policies.

Keywords: Cardiovascular disease. Food consumption. Micronutrients. Machine learning. Clustering Algorithms. Supervised and unsupervised algorithms.

LISTA DE FIGURAS

Artigo 1 - Applicability of machine learning techniques in food intake assessment: A systematic review

Artigo 2 - Cardiovascular disease analysis using unsupervised machine learning approach: Brazilian Cardioprotective Nutritional Program (BALANCE program)

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AIC	*Akaike Information Criterion*
AIT	Ataque Isquêmico Transitório
AGMI	Ácido Graxo Monoinsaturado
AGPI	Ácido Graxo Poliinsaturado
AGS	Ácido Graxo Saturado
AMPM	*Automated Multiple-Pass Method*
AVC	Acidente Vascular Cerebral
AVE	Acidente Vascular Encefálico
BIC	*Bayesian Information Criterion*
cm	Centímetros
CT	Colesterol Total
DAC	Doença Arterial Coronariana
DAP	Doença Arterial Periférica
DCNT	Doenças Crônicas não Transmissíveis
DCV	Doença Cardiovascular
DHA	Ácido Docosahexaenoico
DICA BR	Programa Alimentar Brasileiro Cardioprotetor
dL	Decilitro
DMT2	Diabetes *Mellitus* Tipo 2
DP	Desvio padrão
EPA	Ácido Eicosapentaenóico
g	Grama
HAS	Hipertensão Arterial Sistêmica
HCor	Hospital do Coração
HDL	Lipoproteína de Alta Densidade (do inglês *High Density Lipoprotein)*
IA	Inteligência Artificial
IAM	Infarto Agudo do Miocárdio
IBGE	Instituto Brasileiro de Geografia e Estatística
ICO	Insuficiência Coronariana
IEP	Instituto de Ensino e Pesquisa

IMC	Índice de Massa Corporal
Índice TyG	Índice Triglicerídio-Glicose *(*do inglês*Triglyceride-Glucose Index)*
KDD	*knowledge Discovery in Databases*
Kcal	Quilocalorias
Kg	Quilogramas
Kg/m^2	Quilograma por metro quadrado
L	Litros
LDL	Lipoproteína de Baixa Densidade (do inglês *Low Density Lipoprotein*)
m	Metro
mg	Miligramas
mg/dL	Miligrama por decilitro
ML	*Machine Learning*
MLP	*Multilayer Perceptron*
mmol	Milimol
mmol/L	Milimol por litro
mmHg	Milímetro de mercúrio
NB	*Naïve Bayes*
NCI	*National Cancer Institute*
NIBIB	*National Institute of Biomedical Imaging and Bioengineering*
NIH	*National Institutes of Health*
OMS	Organização Mundial de Saúde
OPAS	Organização Pan-Americana da Saúde
PAD	Pressão Arterial Diastólica
PAS	Pressão Arterial Sistólica
PC	Perímetro da Cintura
PCA	Análise de Componentes Principais *(*do inglês *Principal Component Analysis)*
POF	Pesquisa de Orçamentos Familiares
PROADI-SUS	Programa de Apoio ao Desenvolvimento Institucional do SUS
R24H	Recordatório de 24 horas
RCE	Relação Cintura/Estatura

RI	Resistência a Insulina
SMO	*Sequencial Minimal Optimization*
SUS	Sistema Único de Saúde
SVM	*Support Vector Machine*
TG	Triglicerídeos
VAI	Índice de Adiposidade Visceral (do inglês *Visceral Adiposity Index*)

SUMÁRIO

1 INTRODUÇÃO

As Doenças Cardiovasculares (DCV) pertencem ao grupo das Doenças Crônicas não Transmissíveis (DCNT) com alta incidência em países em desenvolvimento, como o Brasil, e correspondem às principais causas de morte no mundo, sendo responsáveis por cerca de 17,9 milhões destas anualmente (WHO, 2021). No Brasil, 14 milhões de pessoas são acometidas por estas doenças, com mais de 380 mil óbitos todos os anos, sendo o Infarto Agudo do Miocárdio (IAM) responsável por mais de 30% dessas mortes (SBC, 2020). De acordo com previsões, as DCV continuarão sendo a principal causa de morte no mundo até 2030, aumentando para mais de 24 milhões por ano (WORLD HEART FEDERATION, 2019).

As DCV são um grupo de doenças que afetam o coração e os vasos sanguíneos e incluem a doença coronariana, arterial periférica, cerebrovascular, cardiopatia congênita, trombose venosa profunda e embolia pulmonar (WHO, 2021). O aumento da ocorrência dessas doenças está relacionado a diversos fatores de risco, como idade, sexo, história familiar de DCV, predisposição genética, hábito de fumar, sedentarismo, alimentação inadequada, insônia, diabetes *mellitus*, Hipertensão Arterial Sistêmica (HAS), hipercolesterolemia e obesidade (BRASIL, 2020).

No entanto, vale ressaltar que alguns fatores de risco são modificáveis, como a alimentação inadequada. Dietas consideradas inadequadas, ou seja, não saudáveis, são de baixa qualidade, sendo ricas em ácido graxo saturado (AGS), grãos refinados, produtos de origem animal, sódio, doces e bebidas açucaradas, normalmente são produtos alimentícios processados, embalados e prontos para o consumo (SHARIFI-RAD et al., 2020). Um padrão alimentar considerado saudável e que possui efeito cardioprotetor, sendo, portanto, recomendado incluir o consumo de frutas e vegetais, que são alimentos ricos em vitaminas, minerais e fibras, grãos integrais, peixes e alimentos ricos em ácido graxo monoinsaturado (AGMI) e poliinsaturado (AGPI) (NESTEL et al., 2020).

Um dos grandes desafios enfrentados pelas instituições de saúde é o diagnóstico das DCV, pois os métodos mais sofisticados são caros e por isso menos disponíveis. Portanto, geralmente é baseado na anamnese e no exame físico do paciente, contando muitas vezes com a experiência do profissional (PEREIRA et al., 2011). Além disso, outro fator importante é a qualidade na prestação de serviço, que

compreende o diagnóstico precoce e tratamentos seguros e eficazes (MARTINEZ; KING; CAUCHI, 2016).

Motivado pela necessidade de melhorar a qualidade nos serviços de saúde, com redução de custos hospitalares e auxílio na predição de doenças, algoritmos de aprendizagem de máquina, do inglês, *Machine Learning* (ML), tem sido propostos com o objetivo de compreender os grandes conjuntos de dados que são gerados diariamente na área da saúde, auxiliando os profissionais nas tomadas de decisão (MA; CHEN, 2019). O ML é uma subárea da Inteligência Artificial (IA), que tem como objetivo extrair informações ocultas, padrões e dados específicos de grandes bases de dados, a partir do desenvolvimento de algoritmos que concedam a sistemas computacionais a melhoria de sua performance em determinada tarefa. (KODATI; VIVEKANANDAM; RAVI, 2019; SIQUEIRA-BATISTA; SILVA, 2019)

Estes algoritmos vêm sendo aplicados em diversas áreas, como por exemplo, na ciência e nos negócios, bioinformática, setor de compras, entre outros, devido a seu alto desempenho preditivo (KODATI; VIVEKANANDAM; RAVI 2019). Na área da saúde, em que a nutrição está inserida, vêm ganhando destaque ao longo dos últimos anos, uma vez que este setor gera grandes quantidades de dados e tem demonstrado a partir de estudos científicos sua aplicabilidade para avaliação do consumo alimentar, diagnóstico e predição de doenças (TARAWNEH; EMBARAK, 2019).

No entanto, o desenvolvimento e a implementação de algoritmos de ML na área da nutrição ainda é novo, além disso, os profissionais devem ter conhecimento sobre a aplicação e interpretabilidade dos resultados, o que torna bem desafiador. Dessa forma, compreender os aspectos relacionados à escolha e à aplicabilidade dos algoritmos de ML nas DCV e na avaliação do consumo alimentar, e saber interpretar os resultados no ambiente clínico e na pesquisa, poderá ampliar o conhecimento, auxiliando os profissionais na tomada de decisão, com diagnósticos precoces e tratamentos eficazes, e proporcionar aos pesquisadores uma ferramenta precisa e de qualidade para suas investigações.

2 REVISÃO DE LITERATURA

2.1 Panorama das doenças cardiovasculares

As DCV correspondem às principais causas de morte no mundo e, entre elas estão às doenças coronárias, cerebrovasculares, reumáticas, entre outras (WHO, 2021). Estima-se que 17,9 milhões de pessoas morrem a cada ano o que representa 31% de todas as mortes no mundo, com 75% ocorrendo em países de baixa e média renda e um terço em pessoas com menos de 70 anos de idade (WHO, 2020). A Organização Mundial de Saúde (OMS) estabeleceu como meta a redução de 25% das DCNT até o ano de 2025, e em sintonia com a meta para a redução das DCV (WHO, 2017).

Sabe-se que o aumento da ocorrência dessas doenças está relacionado a diversos fatores de risco, como idade, sexo, história familiar de DCV, predisposição genética, hábito de fumar, sedentarismo, alimentação inadequada, distúrbios do sono, diabetes *mellitus*, HAS, hipercolesterolemia e obesidade (BRASIL, 2020). No entanto, vale ressaltar que embora tenham sido identificadas como a principal causa de morte no mundo, são apontadas como as mais evitáveis e controláveis, pois, alguns fatores de risco são modificáveis, como o hábito de fumar, o sedentarismo e a alimentação inadequada (SCHWALM et al., 2016).

É importante destacar que as DCV não são apenas um problema de saúde pública, mas também um desafio econômico. No Brasil, o custo das internações de pacientes com DCV é considerado o maior entre as causas de internações hospitalares, sendo 88% com medicamentos, 66% com previdência social e 33% com morbidade. Estes dados são importantes indicativos de que existe um aumento da população que está convivendo com essas doenças (SIQUEIRA; SIQUEIRA-FILHO; LAND, 2017). Dados do Instituto Brasileiro de Geografia e Estatística (IBGE) mostram que o Brasil está mudando rapidamente a sua estrutura etária, aumentando a proporção de idosos e a expectativa de vida do brasileiro (MIRANDA; MENDES; SILVA, 2016). Sabe-se que o envelhecimento tende a aumentar a incidência de DCV e, consequentemente, os seus custos de forma exponencial (SIQUEIRA; SIQUEIRA-FILHO; LAND, 2017).

No entanto, é preciso ter cuidado ao avaliar apenas as implicações econômicas do tratamento, mas, levar em consideração a capacidade de redução da doença quando o tratamento é pautado na medicina baseada em evidências científicas

(NETO; SILVA, 2008). Para isso, o Ministério da Saúde, guiado pelas diretrizes da OMS e da Organização Pan-Americana da Saúde (OPAS), e com o auxílio de outros ministérios e sociedades científicas, coordenou a elaboração do Plano de Enfrentamento das DCNT no Brasil, no período de 2011 a 2022. O enfoque do Plano gira em torno do estabelecimento de metas para o enfrentamento dos fatores de risco modificáveis como: hábito de fumar; alimentação inadequada; sedentarismo e consumo de bebidas alcoólicas, responsáveis por grande parte da carga de DCNT, dentre elas, as DCV (BRASIL, 2011a).

Embora as DCV tenham sido identificadas como uma das DCNT mais crônicas em todo o mundo é, ao mesmo tempo, a mais evitável. Os dois elementos fundamentais para seu controle são a prevenção primária e secundária. O primeiro aborda a adoção de um estilo de vida mais saudável, com uma alimentação adequada e prática de atividade física, com o objetivo de prevenir a obesidade e outras DCNT. O segundo elemento é focado na prevenção das complicações das DCV, assim como a prevenção de novos eventos (SOCIEDADE BRASILEIRA DE CARDIOLOGIA, 2013). Sendo assim, a triagem precoce e a intervenção médica oportuna podem desempenhar um papel eficiente na prevenção e/ou tratamento destas doenças, a fim de evitar que novos eventos cardiovasculares ocorram (KARUNATHILAKE; GANEGODA, 2018).

Um grande desafio enfrentado pelas organizações de saúde, como hospitais e centros médicos, é a prestação de serviços de qualidade a custos acessíveis. O serviço de qualidade implica diagnosticar precocemente e adequadamente os pacientes e administrar tratamentos eficazes (MARTINEZ; KING; CAUCHI, 2016). Portanto, considerando a necessidade de melhorar a qualidade dos serviços de saúde, reduzir custos hospitalares e investigar a relação do consumo alimentar e as DCV, procedimentos computacionais têm sido propostos principalmente baseados em algoritmos de ML. O objetivo é auxiliar os profissionais no processo de tomada de decisão, além do desenvolvimento e implantação de políticas e sistemas que visem controlar o aumento das DCV (SCHWALM et al., 2016; MA; CHEN, 2019).

2.2 Impacto do consumo alimentar no desenvolvimento das doenças cardiovasculares

O comportamento alimentar sofreu grandes mudanças em todo o mundo, e observa-se um aumento contínuo no consumo de comidas congeladas, carboidratos refinados, bebidas açucaradas, *fast food*, entre outros alimentos ultra processados (PARTULA et al., 2020). Esse padrão alimentar não saudável está associado ao desenvolvimento de DCV e, é caracterizado pelo alto consumo de alimentos de origem animal ricos em AGS, frituras, carnes processadas, sódio, produtos lácteos ricos em gordura, açúcares, bebidas com alto teor de açúcar, consumo excessivo de álcool e alimentos com baixo teor de fibra (MARTINEZ-GONZALEZ et al., 2020).

Evidências demonstram uma associação positiva entre o consumo de carboidratos refinados e de alimentos com alto índice glicêmico com o diabetes *mellitus* tipo 2 (DMT2), obesidade e DCV (GROSS et al., 2004). Já o sódio é um nutriente necessário para uma função fisiológica normal, no entanto, seu consumo elevado aumenta a pressão arterial e a mortalidade por DCV (KOTCHEN; COWLEY; FROHLICH, 2013). Outro alimento que se deve ter cautela em seu consumo são as carnes vermelhas, ricas em proteína, ferro, zinco e vitaminas do complexo B, mas também podem conter quantidades significativas de colesterol e AGS que aumentam a lipoproteína de baixa densidade, do inglês, *low density lipoprotein* (LDL) (MOZAFFARIAN; MICHA; WALLACE, 2010). Além disso, a carne vermelha aumenta a formação endógena de compostos nitrogenados que estão associados com a proliferação epitelial e o estresse oxidativo, por isso seu consumo deve ser equilibrado (BASTIDE; PIERRE; CORPET, 2011).

Nesse sentindo, uma dieta saudável enfatizando o consumo de hortaliças, frutas, legumes, nozes, grãos integrais e peixe são recomendados para diminuir os fatores de risco para DCV. Os óleos vegetais, por exemplo, são ricos em AGPI e reduzem o colesterol total (CT) (ZHENG et al., 2012). O aumento do consumo de laticínios com baixo teor de lipídios está associado a menores concentrações de LDL e triglicerídeos (TG), menor resistência à insulina (RI), perímetro da cintura (PC), índice de massa corporal (IMC), pressão arterial e redução do risco de diabetes *mellitus* (RIDEOUT et al., 2013). Os peixes, por sua vez, são ricos em ácidos graxos ômega-3, ácido docosahexaenóico (DHA) e ácido eicosapentaenóico (EPA),

importantes na redução da inflamação e estresse oxidativo e na manutenção adequada da pressão arterial e da função endotelial (ZHENG et al., 2012).

Estudos demonstraram que o consumo de nozes reduz as concentrações de lipídios séricos, e apesar de sua alta densidade calórica não contribuem ao ganho de peso, talvez devido aos seus efeitos saciadores e aumento das perdas de energia fecal (FLORES-MATEO et al., 2013; SMITH et al., 2015). Já o consumo de leguminosas, frutas e vegetais tem demonstrado redução no PC, CT, pressão arterial, LDL, proteína C reativa e glicose, o que pode reduzir o risco de desenvolvimento de DCV (JAYALATH et al., 2014Ç; LIU et al., 2014).

Evidências mostram ainda redução das DCV com o consumo de uma dieta mediterrânea com alto teor de lipídios totais (MENTE et al., 2009). No ensaio randomizado PREDIMED, uma dieta baseada em vegetais, rica em nozes ou azeite de oliva extra virgem, frutas, legumes, peixes e aves, mas pobre em carnes vermelhas, doces e laticínios integrais foi superior ao grupo controle, atribuído a uma dieta ocidental com baixo teor de lipídios na prevenção de DCV (SHIN et al., 2013). É importante mencionar que as dietas com baixo teor de lipídio não têm sido particularmente eficazes para a redução do risco de DCV em longo prazo, em parte devido à dificuldade de manter tal dieta (LOOK et al., 2013).

2.3 Machine learning e o processo de descoberta do conhecimento

2.3.1 Conceitos importantes sobre Inteligência artificial, mineração de dados e machine learning

O grande volume de dados, conhecido como *big data,* que são gerados em diversas áreas, marca historicamente a era das informações e tem levado a uma mudança nas formas tradicionais de análise de dados (FILHO, 2015). Este aumento expressivo de informações nos bancos de dados gerou grandes desafios para realização dos mesmos, como a escalabilidade, dimensionalidade, qualidade, heterogeneidade, complexidade e distribuição dos dados (TAN; STEINBACH; KUMAR, 2006). Diante disso, foi impulsionado o desenvolvimento de algoritmos e ferramentas computacionais eficientes, com capacidade de potencializar a investigação e transformação dos dados em informações úteis (CARVALHO et al., 2012). Nesse sentido, surgiu o processo denominado como Mineração de Dados, ou do inglês, *Data Mining*, originado por meio da união de técnicas de estatística, IA e ML (TAN; STEINBACH; KUMAR, 2006).

É importante entender o conceito de IA, que pode ser definida como “estudo de agentes que recebem percepções do ambiente e executam ações”. Tais agentes executam suas ações de maneira a maximizar as chances de sucesso para seus objetivos. O campo da IA discute a capacidade das máquinas e dos programas de computador a tomarem decisões com base em dados captados por meio de sensores ou alimentados por meio de intervenção humana e, pode estar associada à tomada de decisão racional em um processo no qual o investigador busca alcançar o melhor resultado (RUSSEL; NORVIG, 2013).

Outro conceito importante é o da Mineração de Dados, que é comumente definida como um processo de caráter multidisciplinar, cujo objetivo é possibilitar a descoberta do conhecimento em um determinado conjunto de dados de forma automática ou semi-automática, por meio de algoritmos que detectam e extraem padrões e informações úteis de forma rápida e precisa (CERQUEIRA et al., 2014).

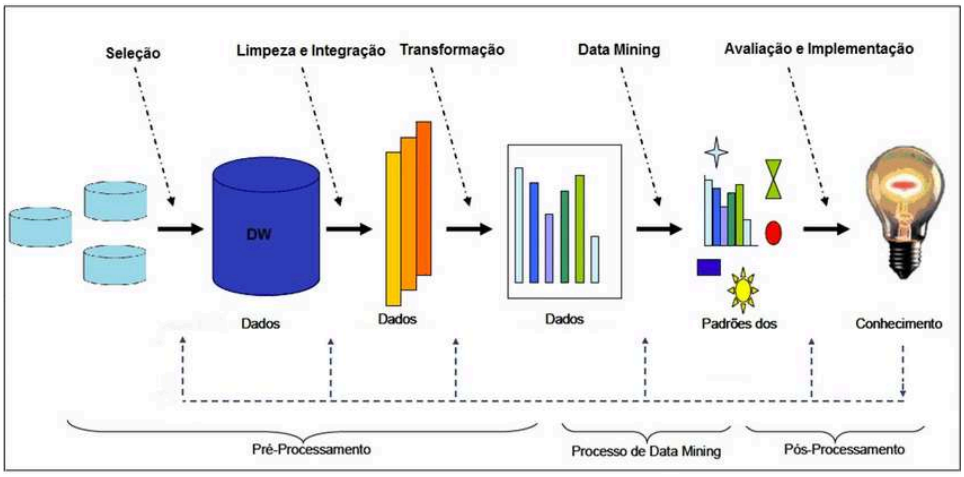
Relacionado aos campos de estudo da IA e mineração de dados, existe outro conceito que muitas vezes se confunde com a mineração de dados e que é denominado ML. Na mineração de dados são utilizadas técnicas para descobrir propriedades e possíveis correlações existentes, podendo então ser utilizados os algoritmos de ML (FERNANDES; FILHO, 2019). O ML é uma subárea da IA que tem como finalidade o estudo e o desenvolvimento de algoritmos que ensinam a sistemas computacionais a desempenhar determinadas tarefas de maneira autônoma. Com o ML é possível reconhecer e extrair padrões de um grande volume de dados, construindo um modelo de aprendizado. Esse aprendizado é baseado na observação de dados, e uma vez que tenham aprendido são capazes de executar tarefas complexas e dinâmicas, prever com precisão, reagir em situações diversas e comportar-se de forma inteligente (KODATI; VIVEKANANDAM; RAVI, 2019; SIQUEIRA-BATISTA; SILVA, 2019).

Atualmente, o estudo de algoritmos de ML tem ganhado destaque devido à alta performance preditiva em análises de grandes volumes de dados. Utiliza-se para análise de compras, no setor de educação, gerenciamento de relacionamento, detecção de fraude e mentira, setor financeiro-bancário, investigação criminal, entre outros (UMASANKAR; THIAGARASU, 2019). Também tem sido utilizado na área da saúde, sendo cada vez mais frequente o uso desses algoritmos para diagnóstico e predição de doenças e, como ferramentas complementares em estudos

epidemiológicos (HORNG et al., 2017; MARUCCI-WELLMAN; CORNS; LEHTO, 2017).

2.3.2 Processo de descoberta do conhecimento

A mineração de dados é uma etapa específica do processo de descoberta de conhecimento em bases de dados (em inglês, *Knowledge Discovery in Databases* [KDD]). Este processo é complexo e interativo e envolve uma sequência de etapas, como visto na Figura 1, sendo elas: pré-processamento, processamento e pós-processamento dos dados (KULKARNI, 2018).



Fonte: SILVA et al (2010).

Figura 1: Etapas da descoberta do conhecimento em banco de dados.

A execução da fase de pré-processamento dos dados é fundamental, uma vez que atributos redundantes ou qualquer outra inconsistência nos dados podem influenciar na detecção de informações relevantes. Portanto, uma condução adequada desta fase pode melhorar o desempenho dos algoritmos (WITTEN; FRANK; HALL, 2016). É nesta etapa onde ocorrem à entrada dos dados brutos, a limpeza, integração, seleção e transformação de dados em formatos adequados para as análises posteriores e, a redução de sua complexidade (KULKARNI, 2018).

Após preparar o conjunto de dados inicia-se a etapa mais duradoura de todo o processo, o processamento, isto é, a fase em que os algoritmos de ML serão aplicados para realizar a tarefa de mineração de dados, com o objetivo de encontrar padrões e informações úteis (TAN; STEINBACH; KUMAR, 2006). Os principais

algoritmos de ML podem ser classificados como de aprendizagem supervisionada e aprendizagem não supervisionada.

Na aplicação dos algoritmos de aprendizagem supervisionada tem-se conhecimento prévio dos valores da variável de saída, ou seja, as classes ou rótulos representados por valores categóricos ou contínuos do conjunto de dados de entrada - composto de registros (instâncias) e variáveis (atributos). O objetivo da aprendizagem supervisionada é aprender empregando algoritmos para este tipo de tarefa, uma função de mapeamento que melhor se aproxime da relação entre os dados de entrada e a saída de modo que, quando novas instâncias estiverem disponíveis, a saída possa ser prevista com precisão (FERNANDES; FILHO, 2019). O conjunto de dados é dividido em duas partes: dados de treinamento e teste. Um modelo preditivo é então construído com base em um algoritmo que usa o conjunto de treinamento para que o modelo resultante aprenda padrões associando os valores dos dados de entrada aos rótulos de saída. Após o treinamento, o modelo receberá a divisão do conjunto de teste, que foi deixada de fora da etapa anterior, e aplicará o conhecimento aprendido com as experiências anteriores (dados de treinamento) a este conjunto de teste para que a precisão, sensibilidade, especificidade - e outras medidas estatísticas importantes - sejam calculadas para avaliar o poder preditivo do modelo (PEDREGOSA et al., 2011).

Os principais algoritmos dessa abordagem são os de classificação e regressão. Os de classificação são usados quando você deseja mapear as variáveis de entrada para uma classe ou categoria específica, ou seja, são utilizados para classificar os dados em uma categoria ou classe, onde a categoria pré-definida é atribuída com base na probabilidade sugerida por um conjunto de treinamento. Os algoritmos baseados em árvores de decisão (*Random Forest*, C4.5), redes neurais (*Multilayer Perceptron* – MLP), *Naïve Bayes* (NB) e *Support Vector Machine* (SVM) (*Sequencial Minimal Optimization* - SMO) são alguns dos métodos de classificação (KHAN et al., 2010). Os modelos de regressão também são utilizados para previsão, mas a principal diferença entre os de classificação e regressão consiste na saída do modelo. Na classificação o resultado são as classes pré-definidas enquanto na regressão o resultado são valores numéricos. De maneira geral, é usado em situações em que o objetivo é mapear as variáveis de entrada para uma saída com um valor contínuo, ou seja, qualquer valor numérico (YILDIZ; BILBAO; SPROUL, 2017). O objetivo da regressão é definir os valores dos parâmetros de uma equação

matemática que define y (a saída a ser prevista) em função das variáveis x (variáveis de entrada). Esta equação, o modelo final, pode então ser usado para prever o resultado para novas instâncias. Em geral, um modelo se ajusta bem aos dados se as diferenças entre os valores observados e os valores previstos forem pequenas e imparciais (PEDREGOSA et al., 2011).

Os algoritmos de aprendizagem não supervisionada são usados para explorar dados não rotulados, ou seja, quando as instâncias não têm valor ou categoria associada. Como resultado, esses algoritmos não visam fazer previsões, mas, em vez disso, encontrar estruturas e padrões ocultos potencialmente úteis que os profissionais possam interpretar e que permitam uma melhor descrição e compreensão dos dados (TAN; STEINBACH; KUMAR, 2006). Nesta abordagem, a tarefa da máquina não é encontrar a saída certa dos dados de entrada, mas explorar os dados e ser capaz de encontrar grupos ou fazer inferências de acordo com as semelhanças, padrões e diferenças encontrados avaliando os recursos das instâncias, sem nenhum treinamento prévio (TAN; STEINBACH; KUMAR, 2006). As tarefas de aprendizagem não supervisionadas normalmente são para encontrar grupos nos dados e/ou revelar regras de associações importantes. Os principais algoritmos são os de agrupamento e os de regras de associação (DEY, 2016).

Nos algoritmos de agrupamento os dados não rotulados são analisados e organizados em grupos por suas semelhanças ou dissimilaridades. A medição de quão semelhantes ou diferentes as instâncias são entre si é feita usando um cálculo de proximidade, como a distância euclidiana (PEDREGOSA et al., 2011). O objetivo é criar um agrupamento (um conjunto de grupos) onde as instâncias no mesmo grupo são muito semelhantes entre si, enquanto as instâncias em grupos distintos são altamente diferentes (ZHENG et al., 2019). Os algoritmos *k*-means, *hierarchical agglomerative clustering*, *expectation-maximization* e *spectral clustering* são os principais exemplos de algoritmos de agrupamento (ZHENG et al., 2019). Já os algoritmos de regras de associação em vez de agrupar instâncias visam descobrir associações ou regularidades potencialmente relevantes entre itens (ou valores de atributos) das instâncias. A seguinte implicação pode representar regras: $X \rightarrow Y$, onde X é chamado de antecedente da regra e Y é chamado de consequente. O algoritmo de regras de associação mais conhecido é o Apriori (GHORBANI; GHOUSI, 2019).

Por fim, a fase de pós-processamento dos dados, responsável pela avaliação, interpretação dos resultados encontrados e apresentação do conhecimento obtido, com o objetivo de ser utilizado para suporte a tomada de decisão no problema estudado (TAN; STEINBACH; KUMAR, 2006). A etapa de avaliação é importante, pois a performance dos modelos obtidos é validada por meio de métodos estatísticos e, a etapa de interpretação é o momento em que o especialista da área verifica os padrões encontrados e define se os resultados foram alcançados (WITTEN; FRANK; HALL, 2016).

2.4 Aplicação dos algoritmos de *machine learning* em estudos sobre doenças cardiovasculares

Os sistemas de saúde em todo o mundo vêm enfrentando muitos desafios, principalmente relacionados à qualidade na prestação de serviços. Além disso, observa-se um aumento da prevalência de DCNT, da morbidade, da transição epidemiológica, da demanda por serviços de saúde, e consequentemente, o aumento dos gastos impactando na economia (ATUN, 2015). A gestão desses sistemas e a qualidade da prestação de serviço são essenciais, e envolve planejamento e recursos como a coleta de informações a partir da triagem de pacientes, exames, investigações, diagnóstico precoce, tratamento eficaz e monitoramento do paciente (PANCH; SZOLOVITS; ATUN, 2018).

Motivados por esses desafios e pela necessidade de se ter um instrumento capaz de extrair as informações geradas diariamente nas grandes bases de dados da área da saúde, algoritmos de ML vêm sendo estudados com objetivos de auxiliar os profissionais da saúde na tomada de decisão e ajudar a compreender as relações existentes entre as características de um indivíduo ou população e determinadas doenças. Ademais, para evitar tanto quanto possível vieses indesejados, erros diagnósticos e custos médicos excessivos, que podem afetar a qualidade do tratamento prestado aos pacientes (MA; CHEN, 2019).

O interesse na aplicação de algoritmos de ML para a área de saúde vem crescendo ao longo dos últimos anos, principalmente pelo seu potencial em auxiliar na interpretação dos dados que são complexos e volumosos, melhorando o desempenho nas decisões de diagnóstico, prognóstico e gestão (BEN-ISRAEL et al., 2020). No entanto, vale ressaltar que seu desenvolvimento e implementação na área da saúde é complexo (CUTILLO et al.,2020), uma vez que devemos levar em

consideração que os profissionais da saúde têm que ter conhecimento sobre a aplicabilidade dos algoritmos e saber interpretar os resultados. Além do mais, é importante que computadores com melhor velocidade de processamento sejam disponibilizados para que esses algoritmos sejam aplicados (STIGLIC et al., 2020).

Ainda assim, nos últimos anos, mais destaque foi dado ao reconhecimento da tomada de decisão na saúde apoiadas pela utilização de algoritmos de ML. Por exemplo, ao prever o risco de doenças, na probabilidade de readmissão do paciente e na previsão da necessidade de cuidados específicos (AHMAD; ECKERT; TEREDESAI, 2018). Por isso, aumentou ainda mais a necessidade de entendimento, compreensão e interpretabilidade (fator chave que limita a adoção mais ampla dos algoritmos de ML na área da saúde) dos processos de ML (STIGLIC et al., 2020).

O *National Institutes of Health* (NIH), em 2019, co-patrocinou um workshop de ML em saúde com o *National Cancer Institute* (NCI), *National Institute of Biomedical Imaging and Bioengineering* (NIBIB), com o objetivo de discutir a aplicabilidade dos algoritmos de ML na área clínica. Destacaram quatro pontos fundamentais, sendo eles: confiabilidade, uma vez que os profissionais da saúde precisam ser capazes de interpretar com precisão e aplicar com segurança as informações derivadas dos algoritmos de ML em um ambiente clínico; explicabilidade, os profissionais têm que ser capazes de compreender e os modelos gerados e, usabilidade, entender até que ponto estes algoritmos podem ser usados para atingir objetivos específicos com eficácia, eficiência e transparência. Ademais, os dados e os algoritmos devem ser disponíveis para todos os profissionais interessados e pessoas que são afetadas por qualquer tipo de decisão, encorajando a transparência dos dados e garantindo acesso aos códigos dos algoritmos (CUTILLO et al.,2020).

Diante disso e do crescimento do uso dos algoritmos de ML na saúde, o poder computacional e a disponibilidade dos bancos de dados vêm melhorando, dessa forma os pesquisadores estão concentrando seus esforços diretamente em tarefas complexas na pesquisa e nos setores de saúde. Esses esforços têm dado excelentes resultados, por exemplo, na radiologia para validação de diagnósticos, na patologia revelando novas características histológicas no câncer de mama e, mais recentemente, na cardiologia para predição de mortalidade por DCV (AL'AREF et al., 2019).

Sendo possível a predição das DCV, as medidas de prevenção e tratamento poderão ser mais eficientes e, maior a possibilidade de redução da mortalidade. Sendo assim, a aplicação dos algoritmos de ML vem trazendo uma nova perspectiva para a previsão dessas doenças (UMASANKAR; THIAGARASU, 2019).

Diferentes algoritmos de ML estão sendo estudados para auxiliar os profissionais de saúde a predizer e diagnosticar as DCV. Os algoritmos mais utilizados são o NB e, os baseados em árvores de decisão e redes neurais (GHORBANI; GHOUSI, 2019). O Quadro 1 apresenta estudos sobre DCV que utilizaram diferentes algoritmos de ML.

Quadro1: Aplicação de diferentes algoritmos de *machine learning* em alguns estudos sobre doenças cardiovasculares.

Autores (ano)	**Objetivo**	**Algoritmos utilizados**	**Conclusão do estudo**
Joshi, Nair (2015)	Predição de DCV	- NB - k-NN - Baseados em árvore de decisão	Não existe um único algoritmo que seja o melhor, sempre há a necessidade de explorar o algoritmo com melhor desempenho no conjunto de dados fornecidos
Chadha et al (2016)	Analisar os diferentes algoritmos de ML que foram propostos nos últimos anos para a predição de DCV	- Baseados em redes neurais - Baseados em árvore de decisão - NB	Os resultados revelam que os algoritmos baseados em redes neurais tiveram um melhor desempenho
Pouriyeh et al (2017)	Investigar a exatidão de diferentes	- Baseados em árvore de decisão - NB	Os resultados indicam que o SVM obteve um melhor desempenho

	algoritmos de ML do tipo classificação na predição das DCV	- MLP - k-NN - SVM	comparado a outros algoritmos
Al-Maqaleh, Abdullah (2017)	Propor um sistema preditivo inteligente utilizando algoritmos de classificação para diagnóstico de DCV	- Baseados em árvore de decisão - Baseados em redes neurais - NB - MLP	Algoritmos aplicados com seleção de atributos superaram os mesmos algoritmos com todos os atributos. A precisão preditiva dos algoritmos é confiável para prever a presença de DCV
Dekamin, Sheibatolhamdi (2017)	Fornecer um método baseado em algoritmos, para diagnóstico de DAC	- *k*-means - NB - k-NN - Baseados em árvores de decisão	Os algoritmos aplicados podem ser utilizados no diagnóstico de DAC
Babu et al (2017)	Avaliar algoritmos de ML não supervisionados e supervisionados para diagnóstico de DCV	- Algoritmo genético - *k*-Means - Baseados em árvores de decisão	Os algoritmos baseados em árvores de decisão têm maior eficiência
Singh, Singh, Pandi-Jai (2018)	Desenvolver um sistema de previsão de DCV utilizando algoritmos de ML	- Baseados em Redes Neurais	O sistema de previsão baseado em redes neurais pode efetivamente predizer DCV

Legenda: DAC – Doença Arterial Coronariana; DCV – Doença Cardiovascular; k-NN - k- Nearest Neighbour; ML – Machine Learning; MLP - Multilayer Perceptron; SVM - Support Vector Machine; NB - NaïveBayes.

Os algoritmos de ML em saúde têm sido aplicados para o diagnóstico de DCV, câncer e diabetes *melittus*. Além disso, podem ser utilizados na prevenção e predição de doenças, diagnóstico precoce, tratamentos eficazes e compreensão dos dados da saúde, além de ajudar a encontrar padrões (GHORBANI; GHOUSI, 2019).

Portanto, como visto acima, a aplicação de algoritmos de ML é uma abordagem eficaz para investigar e analisar grandes conjuntos de dados na área da saúde, como por exemplo, nas DCV. Com o objetivo de descoberta de conhecimento, a fim de auxiliar os profissionais da saúde nas tomadas de decisão, com diagnósticos precoces de doenças e tratamentos eficazes. Essa abordagem também pode auxiliar pesquisadores na elaboração de políticas públicas.

3 OBJETIVOS

3.1 Objetivo geral

Investigar uma população constituída por cardiopatas, na esfera da atenção secundária, segundo suas características sociodemográficas, clínicas, de estilo de vida e de consumo alimentar, utilizando algoritmos de machine Learning.

3.2 Objetivos específicos

- Investigar e compilar os estudos originais que utilizaram abordagem de ML para avaliação do consumo alimentar, de maneira sistemática.
- Realizar uma análise exploratória de dados por meio de algoritmos de ML e métodos de visualização de dados para identificar grupos de perfis multivariados distintos e suas semelhanças e diferenças em pacientes cardiopatas.
- Investigar se o número de eventos cardiovasculares presentes em um indivíduo é influenciado pela ingestão de micronutrientes, e se há diferenças metabólicas e de estilo de vida que impactam também na presença e em seu desenvolvimento.

4 METODOLOGIA

4.1 Delineamento do estudo

Trata-se de um estudo transversal com dados da linha de base do estudo multicêntrico: “Programa Alimentar Brasileiro Cardioprotetor – DICA Br”, registrado em ClinicalTrials.gov (NCT01620398), coordenado por pesquisadores do Instituto de Pesquisa (IEP) do Hospital do Coração (HCor) e viabilizado pelo Programa de Apoio ao Desenvolvimento Institucional do Sistema Único de Saúde (SUS) – PROADI-SUS do Ministério da Saúde.

Fazem parte do DICA Br 34 centros colaboradores das cinco regiões do Brasil, no qual cada centro possui um investigador principal, responsável pela implementação e coordenação do estudo no local e pelo menos dois sub investigadores (WEBER et al., 2016).

O objetivo principal deste programa foi investigar os efeitos do programa alimentar brasileiro cardioprotetor na prevenção secundária de eventos cardiovasculares como parada cardíaca, IAM, acidente vascular cerebral (AVC), revascularização do miocárdio, amputações por Doença Arterial Periférica (DAP), angina ou óbito. Os objetivos secundários foram avaliar a efetividade do estudo na redução de fatores de risco cardiovascular como CT, LDL, glicemia de jejum, pressão arterial, IMC e PC aumentado (WEBER et al., 2016).

4.2 Aspectos éticos

O DICA Br foi aprovado pelo Comitê de Ética do Hospital do Coração (parecer nº 1.171.748) (ANEXO I) e cada centro colaborador submeteu seu protocolo de estudo ao Comitê de Ética local, e os estudos foram iniciados somente após todos os protocolos serem aprovados. O presente trabalho foi incluído como adendo ao projeto DICA Br e aprovado pelo Comitê de Ética em Pesquisa com Seres Humanos da Universidade Federal de Viçosa (pareceres nº 882.612 e 1.020.056) (ANEXO II). O Termo de Consentimento Livre e Esclarecido foi assinado pelos pacientes que aceitaram participar do estudo (WEBER et al., 2016) (ANEXO III).

4.3 Cálculo amostral

O cálculo amostral do estudo DICA Br levou em consideração erro do tipo I de 5%, poder de 80%, taxa de desfecho primário (evento cardiovascular) no grupo controle de 15% e diminuição do risco relativo no grupo intervenção de 30%, o que resultou no tamanho amostral de 2.468 participantes (WEBER et al., 2016). Fizeram parte do presente estudo o total recrutado de 2.535 participantes.

4.4 Recrutamento, critérios de inclusão e exclusão

O recrutamento dos participantes foi realizado mediante parcerias com médicos/residentes, hospitais e centros de referências em tratamento das DCV. Os mesmos foram responsáveis pela identificação dos possíveis participantes e informaram a localização e dados dos pacientes aos investigadores dos centros integrantes do estudo DICA Br para que os pacientes fossem convidados a participar da pesquisa (WEBER et al., 2016).

O período de acompanhamento do DICA Br foi de um mínimo de 36 meses e um máximo de 48 meses, o que compreendeu o período de 38 de março de 2013 a dezembro de 2017. O presente estudo contou com dados da linha de base coletados por pesquisadores treinados durante o período de março de 2013 a dezembro de 2014.

Os critérios de inclusão e exclusão do estudo estão descritos no quadro 2.

Quadro 2: Critérios de inclusão e de exclusão para inserção no DICA Br.

Critérios de Inclusão	**Critérios de Exclusão**
- Idade igual ou superior a 45 anos; - Evidência atual ou nos últimos 10 anos de aterosclerose manifesta, seja ela doença arterial coronariana, doença cerebrovascular ou doença arterial periférica, devidamente confirmada por um médico.	- Presença de condições neuro-cognitivas ou psiquiátricas que dificultem a coleta de dados clínicos confiáveis; - Expectativa de vida inferior a seis meses; - Gravidez ou lactação; - Falência hepática com histórico de encefalopatia ou anasarca; - Insuficiência renal com indicação de

	diálise; - Insuficiência cardíaca congestiva; - Transplante prévio de órgãos; - Uso de cadeira de rodas; - Quaisquer restrições para recebimento de dieta via oral.

Fonte:WEBER et al (2016)

Para a confirmação de doença, adotaram-se os seguintes critérios (WEBER et al., 2016):

a) Doença Arterial Coronariana (DAC)/Insuficiência Coronariana (ICO): presença de um ou mais sintomas:

- DAC assintomática (história de angiografia coronariana ou angiotomografia coronariana com estenose aterosclerótica ≥ 70% do diâmetro de qualquer artéria coronária);
- DAC sintomática (história de angina: diagnóstico clínico, mesmo sem exames complementares; história de positividade a um teste de esforço);
- DAC tratada (realização prévia de angioplastia/*stent*/revascularização);
- IAM (história de IAM) ou síndrome coronariana aguda; história de anormalidade no movimento segmentar da parede cardíaca na ecocardiografia ou um defeito segmentar fixo em cintilografia.

b) Doença cerebrovascular: AVC; Ataque Isquêmico Transitório (AIT); Acidente Vascular Encefálico (AVE), quando o paciente apresentasse um ou mais dos seguintes sintomas:

- Diagnóstico clínico de AVC ou AIT;
- Evidência de AVC prévio na tomografia computadorizada ou na ressonância nuclear magnética.

c) Doença Arterial Periférica (DAP), quando o paciente apresentasse um ou mais dos seguintes sintomas:

- DAP assintomática (relação tornozelo/braço < 0,9 de pressão arterial sistólica em qualquer perna em repouso; estudo angiográfico ou doppler demonstrando estenose > 70% em uma artéria não cardíaca);

- DAP sintomática (claudicação intermitente);
- DAP tratada (cirurgia vascular para doença aterosclerótica);
- Amputação por causa arterial;
- Aneurisma de aorta.

4.5 Desfechos

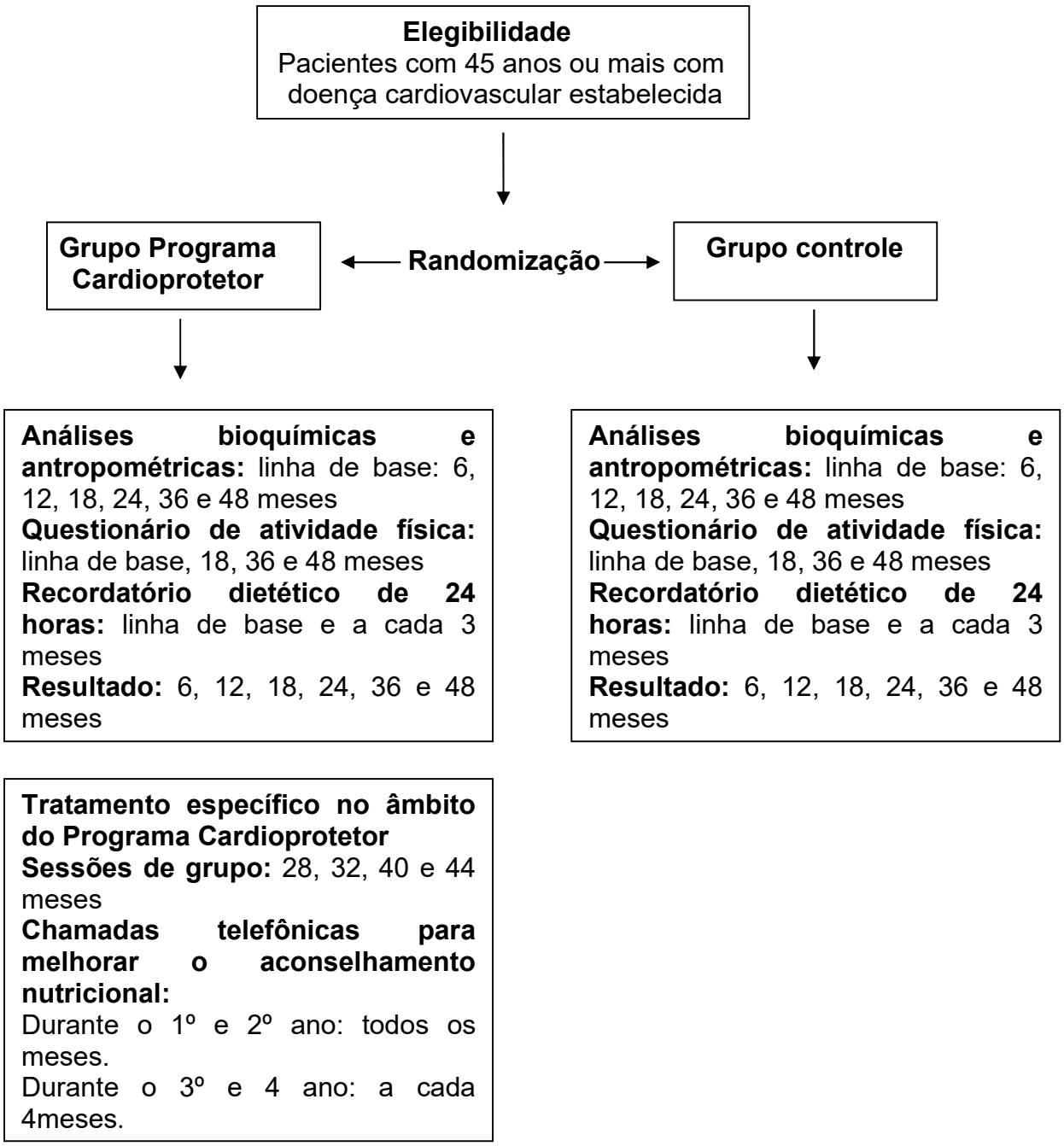
- Acidente Vascular Cerebral
- Amputação de membros
- Aneurisma da aorta
- Infarto Agudo do Miocárdio
- Doença Arterial Coronariana
- Doença Arterial Periférica

4.6 Coleta de dados dos participantes e as variáveis do estudo

Na linha de base do DICA Br foi realizada uma primeira consulta com o paciente, no qual foram obtidas informações quanto às características sociodemográficas, clínicas, uso de medicamentos, prática de atividade física, hábito de fumar, exames bioquímicos, antropometria e consumo alimentar (ANEXO IV). Após 15 dias da inclusão do participante no estudo, foram coletados novamente os dados relacionados ao uso de medicamentos e consumo alimentar (ANEXO V). A razão desta espera foi o cumprimento de um intervalo de tempo após a alta hospitalar dos pacientes que estavam internados (WEBER et al., 2016).

Com exceção dos dados relativos ao consumo alimentar, os demais dados foram digitados em formulário eletrônico (eCRF®). Em relação às informações referentes ao consumo alimentar, foi utilizado o programa computadorizado Nutriquanti® (GALANTE, 2007). Ambos os programas permitem a conversão dos dados em planilhas com formato compatível com versões do *software Microsoft Excel*®.

Apenas as informações da linha de base do estudo fizeram parte da construção desta tese de doutorado. O fluxograma completo da coleta de dados do estudo DICA Br e o fluxograma dos dados e etapas analisados para este estudo estão descritos na Figuras 2 e 3, respectivamente.



Traduzido pela autora (WEBER et al., 2016).

Figura 2: Fluxograma de coleta de dados do estudo DICA Br.

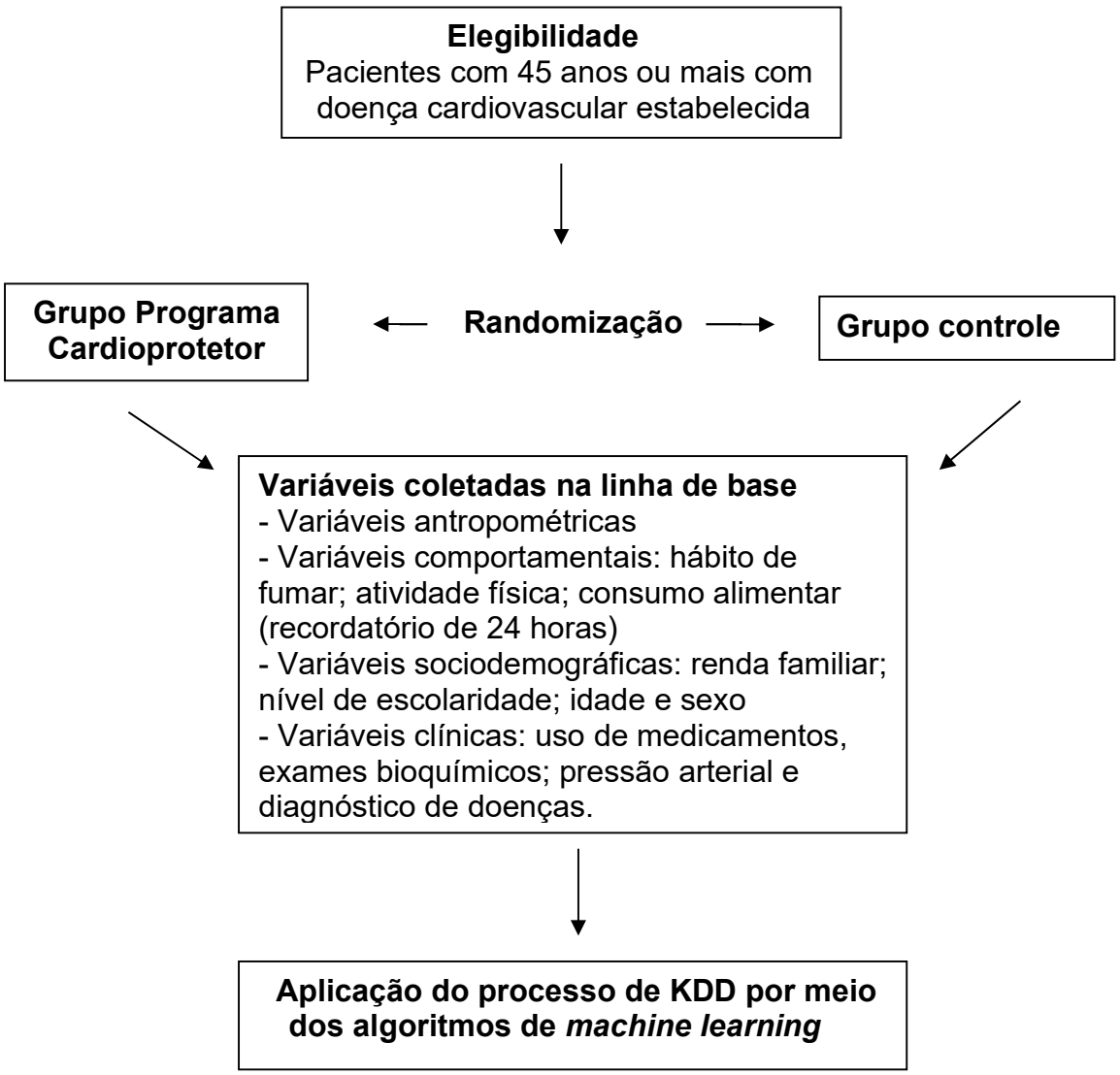


Figura 3: Fluxograma dos dados e etapas analisadas para esta tese.

4.6.1 Variáveis sociodemográficas

Classificação socioeconômica: foi utilizada a classificação da Associação Brasileira de Empresas de Pesquisa para definição da classe econômica do participante, onde ele foi classificado de acordo com as seguintes classes: A1, A2, B1, B2, C1, C2, D e E, para as quais foram levadas em consideração o grau de instrução do chefe de família e a posse, em número de bens de consumo (televisão, rádio, automóvel, empregada mensalista, máquina de lavar, aparelho de DVD, geladeira, freezer e banheiro) (ABEP, 2013).

Escolaridade: foi classificada de acordo com os anos de estudo nos seguintes níveis: analfabeto (nenhum ano de estudo); ensino fundamental (completo ou incompleto); ensino médio (completo ou incompleto); e ensino superior (completo ou incompleto).

Idade: variável discreta. Foram considerados idosos aqueles indivíduos com 60 anos ou mais.

Sexo: os participantes foram classificados como pertencentes ao sexo masculino ou feminino.

4.6.2 Variáveis antropométricas

As variáveis antropométricas foram aferidas por duplas de entrevistadores treinados, utilizando-se a média de cada par de medidas. Fizeram parte da avaliação antropométrica as seguintes medidas:

Peso: utilizou-se balança plataforma mecânica ou digital com precisão mínima de 100 gramas. A medida foi expressa em quilograma (kg);

Estatura: utilizou-se estadiômetro acoplado a balança, portátil, fixo e tipo trena com precisão de 0,5 centímetros, afixados em paredes lisas e sem rodapés. A estatura foi expressa em metros (m);

PC: foi aferido utilizando fita métrica inelástica e flexível, de material resistente. A medida foi realizada na altura do ponto médio entre a borda inferior do arco costal e a crista ilíaca na linha axilar média. (WHO, 2008). O resultado foi expresso em centímetros (cm) e a partir do valor obtido foi determinado o risco de complicações metabólicas, de acordo com os pontos de corte recomendados pela OMS, sendo risco aumentando (≥ 94 cm e ≥ 80 cm) e risco substancialmente aumentado (≥ 102 cm e ≥ 88 cm) para homens e mulheres, respectivamente (WHO, 2008).

4.6.3 Variáveis clínicas

Uso de medicamentos: anti-hipertensivos, anti-coagulantes ou antiplaquetários, hipoglicemiantes e quaisquer outros medicamentos devidamente apresentados com receita médica.

Exames bioquímicos clássicos de risco cardiovascular: as amostras de sangue foram coletadas após jejum de 12 a 14 horas. Marcadores clássicos de risco cardiovascular como TG; CT, glicemia de jejum e Lipoproteína de Alta Densidade, do inglês, *High Density Lipoprotein* (HDL) foram medidos pelo método colorimétrico enzimático (Johnsons & Johnsons, Raritan, EUA, VITROS 5600) e a LDL, determinada pela equação de Friedewald (FRIEDEWALD;LEVY; FREDRICKSON, 1972).

Pressão arterial sistêmica: a Pressão Arterial Sistólica (PAS) e a Pressão Arterial Diastólica (PAD) foram avaliadas por profissionais de saúde treinados, com o paciente em posição confortável, após 5 minutos de repouso com o auxílio de um esfigmomanômetro de mercúrio, seguindo as recomendações da *American Heart Association* (AHA, 2018). Foi classificado como hipertenso quando PAS ≥ 140 ou PAD ≥ 90 mmHg (MALACHIAS et al., 2016) ou uso de medicamentos para o controle da HAS (anti-hipertensivos).

Diagnóstico da DCV: laudo médico indicando a presença de pelo menos um desfecho cardiovascular: DAC na fase assintomática, sintomática ou tratada; DAP na fase assintomática, sintomática ou tratada; AVC, IAM, aneurisma de aorta e amputação por causa arterial.

Diagnóstico de demais doenças relacionadas ao risco cardiovascular: informações auto-referidas sobre antecedentes de HAS, diabetes *mellitus*, dislipidemia e história familiar de DAC.

IMC: calculado a partir do peso (kg)/estatura (m)² para avaliação do estado nutricional de adultos (WHO,1995) e idosos (OPAS, 2002). O IMC foi expresso em quilogramas por metro quadrado (kg/m²) e trabalhado de forma contínua e categórica, para identificar os indivíduos com obesidade.

Relação Cintura/Estatura (RCE): foi calculada pela razão entre o PC (cm) e a estatura (cm) por ser um bom indicador de obesidade central e classificada como alterada quando o resultado foi ≥ 0,5 (ASHWELL; GUNN; GIBSON, 2012).

Índice de Adiposidade Visceral, do inglês, *Visceral Adiposity Index* (VAI): foi calculado pelas fórmulas abaixo para homens (1) e mulheres (2) para estimar disfunções de adiposidade visceral associadas ao risco cardiometabólico (AMATO et al., 2010).

$$\left[\frac{PC\ (cm)}{(39{,}69+1{,}88 * IMC(kg/m2))}\right] * \left(\frac{TG\ (mmol/L)}{1{,}03}\right) * \left(\frac{1{,}31}{HDL\ (mmol/L)}\right) \quad (1)$$

$$\left[\frac{PC\ (cm)}{(36{,}58+1{,}89 * IMC\ (kg/m2))}\right] * \left(\frac{TG\ (mmol/L)}{0{,}81}\right) * \left(\frac{1{,}52}{HDL\ (mmol/L)}\right) \quad (2)$$

RI: foi estimada utilizando o Triglyceride-Glucose Index (Índice TyG), calculado pela fórmula abaixo (3) (SIMENTAL-MENDÍA; RODRÍGUEZ-MORÁN; GUERRERO-ROMERO, 2008; GUERRERO-ROMERO et al., 2010).

$$Ln\left[\frac{TG\ em\ jejum\ (mg/dl) * glicemia\ em\ jejum\ (mg/dl)}{2}\right] \quad (3)$$

Fatores de risco cardiometabólicos: obesidade; inatividade física; hábito de fumar; PC elevado; HAS; e os marcadores de risco tradicionais, como altas concentrações de CT, LDL, TG e glicemia de jejum, e baixas concentrações de HDL, classificados de acordo com a Diretriz Brasileira de Dislipidemias e Prevenção da Aterosclerose (FALUDI et al., 2017).

4.6.4 Variáveis comportamentais

Hábito de fumar: fumante, ex-fumante ou nunca fumou (WEBER et al., 2016).

Exposição ao fumo: nunca, 1 vez por semana, 2 vezes por semana, 3 a 6 vezes por semana e diariamente (WEBER et al., 2016).

Atividade física: avaliada no âmbito do lazer. Os critérios para classificação da prática de atividade física foram (HASKELL et al., 2007; WEBER et al., 2016):

Sedentário: indivíduo que não realizou nenhuma atividade física por pelo menos 10 minutos contínuos durante a semana;

Insuficientemente ativo: indivíduo que realiza atividades físicas, porém de forma insuficiente para ser classificado como ativo, pois não cumpre as recomendações quanto à frequência ou duração;

Ativo: indivíduo que praticou atividade vigorosa em 3 dias ou mais na semana por 20 minutos ou mais por sessão ou atividade moderada ou caminhada em 5 dias ou mais na semana por 30 minutos ou mais por sessão ou qualquer atividade somada que resulte em 5 ou mais dias na semana por 150 minutos por semana;

Muito ativo: indivíduo que pratica atividade vigorosa em 5 ou mais dias na semana por 30 minutos ou mais por sessão ou atividade vigorosa em 3 ou mais dias da semana por 20 minutos ou mais + atividade moderada em 5 ou mais dias da semana por 30 minutos por sessão.

Avaliação do consumo alimentar

O inquérito alimentar utilizado foi o Recordatório de 24 Horas (R24H), o qual permitiu obter informações do participante a respeito do tipo de alimento e/ou preparações consumidas, tamanho das porções em medidas caseiras e/ou gramas, marcas comerciais e os horários nos quais as refeições foram consumidas nas 24 horas anteriores à consulta. Os dados referentes ao consumo alimentar foram

coletados na visita clínica inicial e na visita clínica de 15 dias. Foi respeitado um prazo máximo de 15 dias entre a coleta do primeiro R24H e o segundo para posteriormente realizar a média entre os dois. Os dados foram provenientes de todos os sete dias da semana, podendo ambos R24H coletados serem referentes a dias de semana.

Todos os pesquisadores de campo envolvidos com a coleta de dados foram capacitados pelos membros da coordenação do estudo DICA Br. O treinamento envolveu uma etapa teórica, onde houve a demonstração de um vídeo a respeito da metodologia e aplicação do R24H e outra parte prática, na qual, houve aplicação do R24H entre os pesquisadores.

O método utilizado para a aplicação do R24H teve como referência o *Automated Multiple-Pass Method* (AMPM) que consistiu em uma entrevista guiada por cinco etapas, onde: 1) obteve-se do entrevistado uma lista de alimentos consumidos durante todo o dia; 2) perguntou-se a respeito de alimentos possivelmente esquecidos (ex. balas, petiscos, refrigerantes, etc); 3) obteve-se informações sobre os horários e locais da realização das refeições; 4) fez-se um detalhamento dos alimentos relatados e; 5) revisou-se os alimentos potencialmente esquecidos.

No entanto, a técnica que foi utilizada para se obter as informações do entrevistado sofreu modificações na primeira etapa. Optou-se por questionar sobre os alimentos consumidos pelo entrevistado por período do dia, ao invés de se obter inicialmente uma lista de todos os alimentos consumidos durante todo o dia.

O estudo DICA Br contou com a elaboração de um álbum fotográfico para padronização de medidas caseiras e porções de alimentos. Nele, foram ilustradas porções de alimentos e medidas caseiras relatadas pela população brasileira na Pesquisa de Orçamentos Familiares (POF) do IBGE (BRASIL, 2011b).

Foi utilizado o programa computadorizado Nutriquanti® (GALANTE, 2007) o qual prioriza a utilização de tabelas de composição de alimentos brasileiras e americanas, sendo elas: Tabela Brasileira de Composição de Alimentos (TACO, 2006); *Table of Nutrient Retention* (USDA, 2003); Tabela de Composição de Alimentos (PHILLIPI, 2001) e Tabela de Composição dos Alimentos (BRASIL, 2011b) para o cálculo dos macronutrientes e micronutrientes, sendo eles: calorias, carboidrato, proteína, lipídio, AGS, AGI, AGPI, ácido graxo trans, colesterol, sódio, fibra, cálcio, ferro, potássio, magnésio, fósforo, cobre, zinco, selênio, ômega 3,

ômega 6, vitaminas (A, B1, B2, B3, B5, B6, B7, B12, C, D e E). Cada nutriente foi ajustado por 1.000 kcal de ingestão de energia.

Com base nas informações fornecidas pelo R24H, os alimentos e preparações consumidos pelos participantes foram classificados por meio da *NOVA*. A *NOVA* é uma classificação brasileira que agrupa os alimentos em quatro grupos (alimentos *in natura* ou minimamente processados; ingredientes culinários processados; alimentos processados e alimentos ultra processados) de acordo com a extensão e o propósito do processamento aos quais esses alimentos são submetidos (MONTEIRO et al., 2016). Fizeram parte destes grupos, os seguintes alimentos:

Alimentos in natura ou minimamente processados: legumes; verduras; frutas; raízes e tubérculos; cereais; nozes; frutas secas; ovos, carnes, aves, peixes e frutos do mar; entre outros;

Ingredientes culinários processados: açúcar; mel; óleos e gorduras de origem vegetal ou animal; amidos; sal; vinagre; bebidas alcoólicas em preparações; entre outros;

Alimentos processados: conservas em geral; castanhas salgadas ou doces; carnes salgadas; queijos; pães; entre outros;

Alimentos ultraprocessados: refrigerantes; sucos em pó; sorvetes; chocolates; balas e guloseimas em geral; biscoito recheado, bolos; cereais matinais; caldos liofilizados; salsicha; hambúrguer e outros produtos de carne reconstituída; entre outros.

As preparações mistas foram o maior desafio para classificação segundo a *NOVA.* Pensando nisso, essas preparações foram classificadas de acordo com a proporção dos ingredientes principais utilizados. Dessa forma, uma vez que a maior proporção dos ingredientes principais foi proveniente de alimentos ultraprocessados, esta preparação foi considerada como integrante do grupo de alimentos ultraprocessados e o mesmo aconteceu para os demais grupos de alimentos.

O perfil de consumo alimentar foi expresso de acordo com a contribuição percentual das calorias fornecidas pelos alimentos agrupados pela classificação *NOVA* em relação à ingestão energética diária.

4.7 Controle de qualidade dos dados

A qualidade dos dados foi garantida pela introdução de dados automatizada, contato mensal com os investigadores, visitas de monitoramento aos centros e monitoramento estatístico central. O *feedback* aos investigadores foi fornecido mediante reuniões e *newsletters* periódicos.

Com o intuito de se obter a consistência dos dados digitados e minimizar erros por sub ou superestimação do consumo alimentar, os pesquisadores envolvidos com a coleta de dados foram capacitados constantemente. Além disso, foram selecionados aleatoriamente 20% dos R24H processados, para se avaliar a consistência dos dados.

Quando necessário, foram solicitadas correções e os erros mais comuns utilizados como recurso para reforço nas capacitações. Os principais tipos de erros avaliados foram os de digitação de medida caseira, a omissão ou duplicidade do alimento ou falta de detalhamento das informações coletadas.

4.8 Metodologia da aplicação dos algoritmos de machine learning no banco de dados do DICA Br

4.8.1 Ferramentas computacionais

Para o desenvolvimento de todo o processo de KDD, assim como a aplicação dos algoritmos de ML na coleta de dados, foram empregadas ferramentas computacionais para auxiliar em todas as etapas da descoberta do conhecimento. Nesse sentindo, foi utilizado o ambiente de desenvolvimento Jupyter Notebook (versão 6.0.3) (KLUYVER et al., 2016) em conjunto com a linguagem de programação Python (versão 3.7.6) (PEDREGOSA et al., 2011). Para emprego dos recursos computacionais de análise dos dados, foram importadas do Jupyter Notebook as bibliotecas de ciência de dados, como por exemplo, pandas e numpy para manipulação de dados, matplotlib para visualização de dados, scipy para análises estatísticas e sklearn para os algoritmos de ML.

Para o aprimoramento da análise visual dos resultados sempre que necessário foi utilizada a ferramenta webColorBrewer 2.0 (https://colorbrewer2.org/) para escolher a melhor paleta de cores de forma que indivíduos daltônicos possam visualizar.

Os experimentos computacionais foram realizados em um notebook comum com CPU de 8GB RAM, processador Intel Core i7-4500U 1,80 GHz, sob o sistema operacional Windows 8.1 Pro 64-bit.

4.8.2 Pré-processamento dos dados (1ª etapa)

Esta fase caracteriza-se pela exclusão de dados redundantes ou qualquer outra inconsistência que podem influenciar a detecção de informações relevantes. Portanto, a realização desta fase pode melhorar o desempenho dos algoritmos de ML (WITTEN; FRANK; HALL, 2016).

Nesta fase foram realizadas as seguintes etapas

a) *Limpeza dos dados:* dados inconsistentes, nulos e *outliers*.

b) *Integração dos dados:* reunião dos dados de diferentes fontes a fim de produzir informações relevantes e proporcionar uma visão mais consistente.

c) *Seleção de atributos:* os dados relevantes para a aplicação dos algoritmos de ML foram identificados e reunidos, formando um subconjunto do banco de dados, descartando algum resquício de dados irrelevantes.

d) *Transformação de dados:* transformação ou consolidação dos dados em forma apropriada para a aplicação dos algoritmos.

Essas etapas são extremamente relevantes para simplificar o modelo sem perder informações importantes. Além disso, a etapa de seleção dos atributos reduz a complexidade da base de dados, o que torna a construção do modelo mais rápida e, em muitos casos, produz melhores resultados (YANG; PEDERSEN, 1997).

As quatro etapas citadas acima foram realizadas em conjunto pela equipe de pesquisa. A retirada manual dos atributos (variáveis) e das instâncias (pacientes) foi baseada na redundância e nos valores ausentes que podem comprometer a qualidade dos dados. Nesta fase foram retiradas as seguintes relações de atributos: atributos dos grupos alimentares, percentual de carboidratos, proteína e lipídios, e atributos categóricos, sendo eles: obesidade, dislipidemia, história de DAC, HAS e diabetes *mellitus,* hábito de fumar, atividade física, frequência de exposição ao fumo, e os dez eventos cardiovasculares

Além disso, foi realizada também a análise descritiva dos dados com base na média, desvio padrão, valores extremos, percentis e quartis. Em seguida, a dispersão, variância e outliers foram analisadas graficamente usando histogramas e boxplots. Essas análises têm como objetivo confirmar se todos os atributos e

instâncias que permaneceram após a retirada manual são realmente importantes e para se ter melhor visualização e compreensão inicial das características básicas e fundamentais dos dados.

Ainda, foi realizada uma etapa fundamental, a normalização dos atributos. Os valores dos atributos numéricos foram normalizados para o intervalo [0, 1], enquanto que os valores dos atributos nominais foram mapeados para valores ordinais entre 0 e 1. Por exemplo, os valores dos atributos binários foram mapeados para 0 e 1, os valores dos atributos ternários foram mapeados para 0, 0,5, 1 e assim por diante.

Com o objetivo de diminuir ainda mais o número de atributos, utilizamos um algoritmo de redução de dimensionalidade, conhecido como Análise de Componentes Principais, no inglês *"Principal Component Analysis"* (PCA) (DING; HE, 2004). Sabe-se que a diminuição da dimensionalidade reduz a complexidade dos dados, o que muitas vezes faz com que os algoritmos rodem mais rapidamente e produzam melhores resultados (CERQUEIRA et al., 2014). O PCA é uma transformação linear ortogonal dos dados para um novo sistema de coordenadas de forma que a maior variância se encontre na primeira coordenada (chamada de primeiro componente principal), a segunda maior variância encontra-se na segunda coordenada (chamada de segundo componente principal), e assim por diante. Portanto, a fim de reduzir a dimensionalidade, basta eliminar os componentes de classificação inferior (atributos recém-criados), ou seja, manter apenas os componentes que capturam quantidade suficiente de variação nos dados (CERQUEIRA et al., 2014).

4.8.3 Processamento dos dados (2ª etapa)

Fase em que são executados os algoritmos de ML sobre o conjunto de dados já padronizado para análise dos dados e reconhecimento de padrões e *insights*. A escolha dos algoritmos a serem aplicados depende do objetivo que se deseja alcançar. Os principais algoritmos de ML são os de aprendizagem supervisionada e não supervisionada (VILLACAMPA, 2015).

Depois de preparar o conjunto de dados, os algoritmos de ML foram aplicados para realizar a tarefa de mineração de dados. No caso deste estudo, não houve conhecimento prévio sobre correlações, padrões ou tendências no conjunto de dados obtido, ou seja, não existiu uma classe de interesse definida previamente. Portanto, optou-se por uma abordagem de ML não supervisionada, mais

especificamente, a aplicação de algoritmos de agrupamento para realizar uma análise exploratória dos dados resultantes.

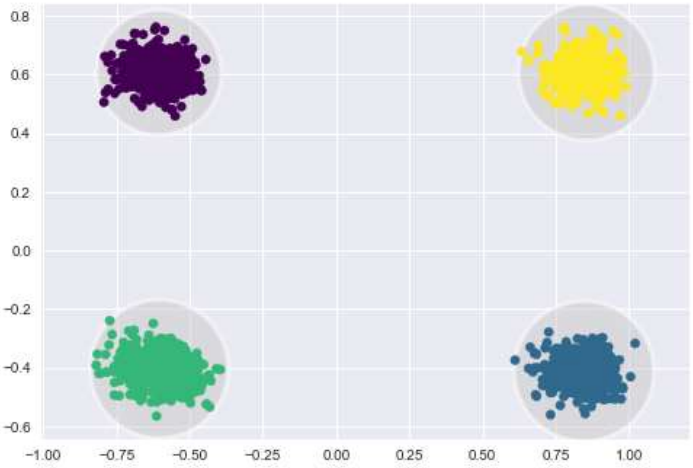
4.8.3.1 Aplicação dos algoritmos de abordagem não supervisionada

Uma das mais frequentes tarefas na abordagem não supervisionada e que fazem parte do presente estudo, dizem respeito a agrupamento. Os algoritmos de agrupamento são usados para explorar dados não rotulados, ou seja, conjuntos de dados cujas instâncias não estão previamente associadas a qualquer categoria (ou classe). O objetivo é encontrar grupos coesos e bem separados - que podem ser interpretados como categorias ocultas - e depois estudar cada grupo separadamente para revelar informações anteriormente desconhecidas (GHORBANI; GHOUSI, 2019).

Uma etapa fundamental para estabelecer os grupos é o cálculo da proximidade entre as instâncias, ou seja, medir o quão semelhantes ou diferentes as instâncias são entre si. Existem muitas alternativas para realizar esse cálculo, dependendo principalmente do tipo de atributos. Um método bastante conhecido e geral é a distância euclidiana (ZHENG et al., 2019), que foi a medida de proximidade escolhida em nosso trabalho.

Os seguintes algoritmos de agrupamento foram utilizados:

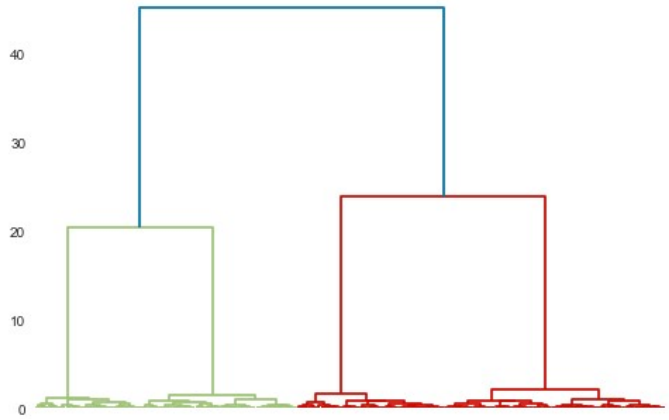
K-means: constrói um agrupamento, sendo o número de grupos definido pelo parâmetro *k*. Este algoritmo distribui as instâncias entre k grupos de acordo com a proximidade, i.e., para cada instância i mede-se a distância entre i e o centro de cada grupo. A instância i será associada ao grupo que resultar na menor distância (JAIN, 2010). Os principais passos do algoritmo são: i) inicialização aleatória dos centróides, também conhecidos como centro dos grupos; ii) cálculo da distância entre cada instância do conjunto de dados e os centróides; iii) associação da instância ao grupo com menor distância; iv) após o agrupamento de todas as instâncias dos dados é realizada a atualização dos centróides por meio do cálculo da média das instâncias associadas a cada centróide. Esse passo faz com que cada centróide que foi inicializado de maneira aleatória seja ajustado para o correto centro do grupo; v) repetição dos passos ii, iii e iv até que não haja mais nenhuma modificação no agrupamento dos dados e nem a necessidade de se atualizar os centróides (JAIN, 2010). Após a execução desses passos, o algoritmo é finalizado e os k grupos finais são dados como saída (JAIN, 2010). Ilustrado na figura 4.



Fonte: Autoria própria.

Figura 4: Ilustração do agrupamento do algoritmo k-means.

Hierarchical clustering: Os agrupamentos são obtidos pela representação dos grupos em dendograma que consiste de um tipo especial de árvore, na qual os nós pais agrupam os nós filhos. Dessa forma, o *hierarchical clustering* agrupa os dados de modo que se dois exemplos são agrupados em algum nível, nos níveis mais acima eles continuam fazendo parte do mesmo grupo, construindo uma hierarquia de grupos. Esse algoritmo permite analisar os grupos em diferentes níveis, pois cada nível do dendograma descreve um conjunto diferente de agrupamentos (BRUM; MOZZAQUATRO; ZANATTA, 2019). Ilustrado na figura 5.

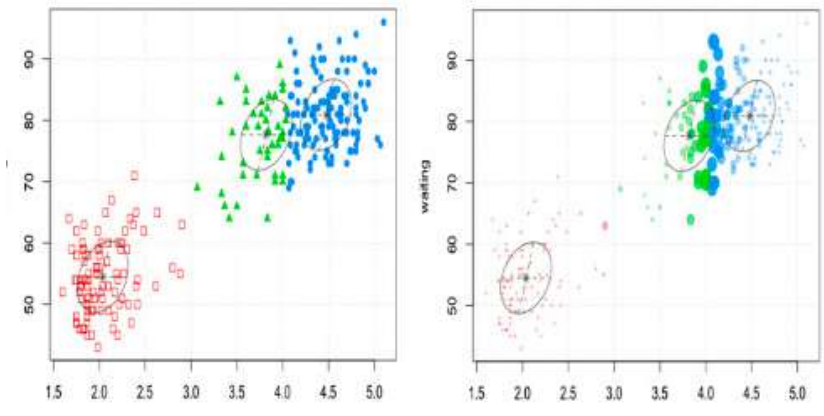


Fonte: Autoria própria.

Figura 5: Dendograma ilustrativo do algoritmo Hierarchical clustering.

Expectation-maximization algorithm: é um algoritmo iterativo usado para encontrar parâmetros de máxima verossimilhança. Este algoritmo alterna entre executar a etapa de expectativa e a de maximização. A etapa de expectativa cria uma função para a expectativa da verossimilhança logarítmica usando a estimativa

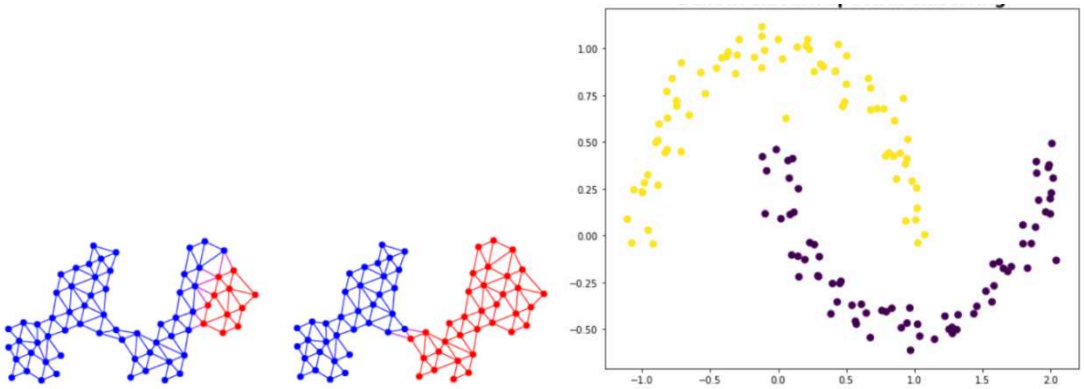
atual para os parâmetros. A etapa de maximização calcula parâmetros para maximizar a verossimilhança logarítmica encontrada na etapa de expectativa. Essas estimativas de parâmetro são usadas para determinar a distribuição das variáveis latentes na próxima etapa de expectativa, e o algoritmo se repete várias vezes, por isso, é chamado iterativo (JUNG; KANG; HEO, 2014). Ilustrado na figura 6.



Fonte: SERRA; TAGLIAFERRI (2019).

Figura 6: Ilustração do algoritmo Expectation Maximization.

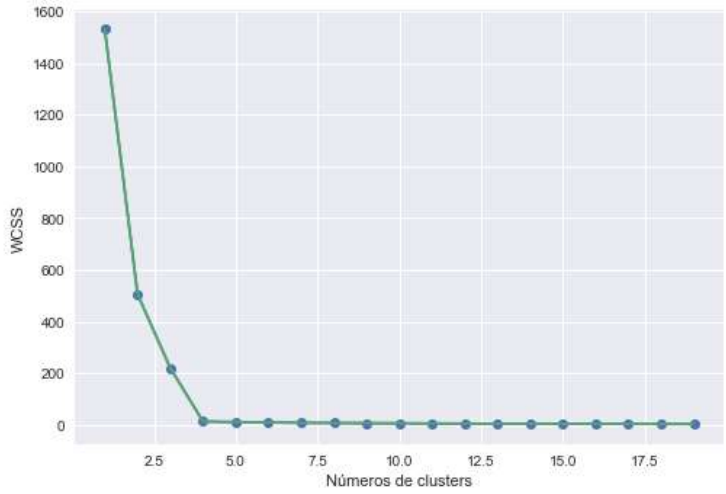
Spectral clustering: vem da teoria da partição do gráfico spectral. O desempenho deste algoritmo está intimamente relacionado ao critério de partição do gráfico correspondente, que visa encontrar uma partição gráfica mais equilibrada e ideal. A principal tarefa do método spectral é obter informações da estrutura organizacional dos dados a partir da relação de similaridade entre eles, de modo que os agrupamentos aconteçam de forma mais natural do que outros algoritmos como, por exemplo, o *k*-means. Para uma melhor compreensão do método, considere um conjunto com n pontos $x_1, ..., x_n$. Cada par de pontos é relacionado por um parâmetro de similaridade levando em consideração a Distância Euclidiana $f(d(x_i , x_j), \theta)$, e assim, é construída uma matriz simétrica e positiva com as relações de similaridades entre cada par (GARCIA; EMMENDORFER, 2017). Ilustrado na figura 7.



Fonte: Nahshon (2018).

Figura 7: A ilustração apresenta os pontos de dados conectados ao gráfico (esquerda) e posteriormente o gráfico do algoritmo Spectral clustering finalizado (direita).

Antes da aplicação de cada algoritmo é importante definir o número de grupos que esses algoritmos precisam produzir. Uma abordagem simples e eficaz é testar vários valores e calcular algumas medidas de qualidade - como o coeficiente de silhueta - para cada agrupamento resultante, de modo que todos os agrupamentos obtidos possam ser comparados. Com base nessa idéia, o método do cotovelo ou método Elbow Curve, ilustrado na figura 8, é uma métrica usada para encontrar a quantidade ideal de grupos *k*. Este método testa a variância dos dados em relação ao número de grupos e foi aplicado no caso do *K*-means (JAIN, 2010), *hierarchical clustering* (BRUM; MOZZAQUATRO; ZANATTA, 2019) e *spectral clustering* (DING et al., 2018). Para o *expectation-maximization algorithm*, utilizamos a convergência dos métodos estatísticos *Akaike Information Criterion* (AIC) e *Bayesian Information Criterion* (BIC)(ZHENG et al., 2019).



Fonte: Autoria própria.

Legenda: WCSS – Elbow method is used within cluster sumo f errors.

Figura 8: Ilustração do método do cotovelo.

Para comparar a qualidade dos agrupamentos de cada algoritmo testado utilizamos o gráfico de dispersão e o coeficiente de silhueta, que combina medidas de coesão e separação, ou seja, seu valor denota quão intimamente relacionadas estão às instâncias dentro de um grupo e quão bem separados os grupos estão uns dos outros (JAIN, 2010; PEDREGOSA et al., 2011). O coeficiente de silhueta varia no intervalo [-1, 1]. Valores próximos de 1 indicam um agrupamento de alta qualidade, ou seja, coesão e separação satisfatórias, valores próximos a zero significam agrupamento indiferentes, ou seja, a distância entre os grupos não é significativa, e valores negativos indicam que um número de instâncias pode ter sido atribuído ao grupo errado (JAIN, 2010; PEDREGOSA et al., 2011; AHMAD; ECKERT; TEREDESAI, 2018).

4.8.4 Pós-processamento dos dados (3ª etapa)

A última etapa do procedimento de análise dos dados é a avaliação dos padrões, apresentação e assimilação dos conhecimentos obtidos. Apresentamos os resultados de forma visual e legível. Neste sentindo, os resultados foram representados, analisados e interpretados usando gráficos de paralelo de coordenadas, gráficos de radar e boxplot. Os gráficos de paralelo de coordenadas e de radar permitiram uma análise de dados multivariados em alta dimensionalidade ao passo que o boxplot permitiu uma análise univariada (LI; ZHEN; YAO, 2017).

4.9 Análises estatísticas

4.9.1 Análise estatística do artigo 2

Após a definição dos quatro grupos encontrados por meio dos algoritmos de ML não supervisionados, a última etapa do procedimento de análise dos dados foi o estudo e comparação desses grupos resultantes visando à extração de padrões e conhecimento. Como uma primeira abordagem, os resultados foram representados, analisados e interpretados usando gráficos de coordenadas paralelas. Esses gráficos representam dados socioeconômicos, clínicos, bioquímicos, de estilo de vida e eventos cardiovasculares, bem como os fatores de risco para DCV. Um gráfico paralelo de coordenadas permite a análise de dados multivariados em alta dimensionalidade desenhando n linhas para n grupos. As linhas representam os valores médios de cada atributo em cada grupo e permitem a comparação dos valores dos atributos em grupos distintos. Os valores do eixo y são normalizados em uma escala de 0 a 1.

Os testes estatísticos clássicos foram aplicados em uma segunda abordagem para comparar o conteúdo dos grupos obtidos. Primeiramente, foi aplicado o teste de Shapiro-Wilk para testar a normalidade dos atributos, e observou-se que eles não apresentavam distribuição normal. Portanto, as comparações estatísticas dos atributos entre os grupos foram realizadas usando testes não paramétricos. Para a comparação dos quatro grupos foi utilizado o teste de Kruskal-Wallis, seguido do teste post-hoc de Conover. Observe que, neste caso, a mediana (intervalo interquartil) é usada nos testes em vez da média. Os atributos categóricos, por sua vez, foram avaliados pelos testes Qui-quadrado de Pearson e Exato de Fisher.

Para a aplicação dos algoritmos de ML foi utilizado o ambiente de desenvolvimento Jupyter Notebook (versão 6.0.3) (KLUYVER et al., 2016) em conjunto com a linguagem de programação Python (versão 3.7.6) (PEDREGOSA et al., 2011). Foram importadas do Jupyter Notebook as bibliotecas de ciência de dados, como o pandas e numpy para manipulação de dados, matplotlib para visualização de dados, scipy para análises estatísticas e sklearn para os algoritmos de ML.

As análises estatísticas foram conduzidas com auxílio do Python versão 3.7.6 admitindo um nível de confiança (α) de 5% para todas as análises realizadas.

4.9.2 Análise estatística do artigo 3

Os indivíduos foram classificados em dois grupos de acordo com o número de eventos cardiovasculares (1 evento ou ≥ 2 eventos). A normalidade das variáveis foi avaliada por meio do teste de Shapiro-Wilk, sendo constatado que os dados não seguem uma distribuição paramétrica. Dessa forma, os resultados são apresentados em mediana (intervalo interquartil). Para a comparação estatística entre os dois grupos foi realizado o teste não paramétrico de Mann–Whitney. O teste qui-quadrado de Pearson foi realizado para avaliar associações potenciais entre as características de estilo de vida (atividade física e hábito de fumar) e os dados sociodemográficos.

As análises estatísticas foram realizadas utilizando a linguagem de programação Python (versão 3.7) por meio das bibliotecas numpy e scipy.stats para manipulação de dados e estatística, respectivamente.

5 RESULTADOS

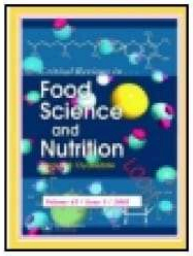
Os resultados desta tese estão estruturados sob a forma de três artigos científicos, sendo:

Artigo 1: Applicability of machine learning techniques in food intake assessment: A systematic review.

Artigo 2: Cardiovascular disease analysis using unsupervised machine learning approach: Brazilian Cardioprotective Nutritional Program (BALANCE program).

Artigo 3: Baixa ingestão de micronutrients está associada com eventos cardiovasculares em pacientes em atenção secundária

5.1 Artigo 1: Applicability of machine learning techniques in food intake assessment: A systematic review.



Taylor & Francis
Taylor & Francis Group

Critical Reviews in Food Science and Nutrition

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/bfsn20

Applicability of machine learning techniques in food intake assessment: A systematic review

Larissa Oliveira Chaves, Ana Luiza Gomes Domingos, Daniel Louzada Fernandes, Fabio Ribeiro Cerqueira, Rodrigo Siqueira-Batista & Josefina Bressan

To cite this article: Larissa Oliveira Chaves, Ana Luiza Gomes Domingos, Daniel Louzada Fernandes, Fabio Ribeiro Cerqueira, Rodrigo Siqueira-Batista & Josefina Bressan (2021): Applicability of machine learning techniques in food intake assessment: A systematic review, Critical Reviews in Food Science and Nutrition, DOI: 10.1080/10408398.2021.1956425

To link to this article: https://doi.org/10.1080/10408398.2021.1956425

Published online: 29 Jul 2021.

CRITICAL REVIEWS IN FOOD SCIENCE AND NUTRITION
https://doi.org/10.1080/10408398.2021.1956425

Taylor & Francis
Taylor & Francis Group

REVIEW

Applicability of machine learning techniques in food intake assessment: A systematic review

Larissa Oliveira Chaves[a], Ana Luiza Gomes Domingos[a], Daniel Louzada Fernandes[b], Fabio Ribeiro Cerqueira[c], Rodrigo Siqueira-Batista[d,e], and Josefina Bressan[a]

[a]Department of Nutrition and Health, Universidade Federal de Viçosa, Viçosa, Brazil; [b]Department of Informatics, Universidade Federal de Viçosa, Viçosa, Brazil; [c]Department of Production Engineering, Universidade Federal Fluminense, Petrópolis, Brazil; [d]Department of Medicine and Nursing, Universidade Federal de Viçosa, Viçosa, Brazil; [e]School of Medicine of the Faculdade Dinâmica do Vale do Piranga, Ponte Nova, Brazil

ABSTRACT

The evaluation of food intake is important in scientific research and clinical practice to understand the relationship between diet and health conditions of an individual or a population. Large volumes of data are generated daily in the health sector. In this sense, Artificial Intelligence (AI) tools have been increasingly used, for example, the application of Machine Learning (ML) algorithms to extract useful information, find patterns, and predict diseases. This systematic review aimed to identify studies that used ML algorithms to assess food intake in different populations. A literature search was conducted using five electronic databases, and 36 studies met all criteria and were included. According to the results, there has been a growing interest in the use of ML algorithms in the area of nutrition in recent years. Also, supervised learning algorithms were the most used, and the most widely used method of nutritional assessment was the food frequency questionnaire. We observed a trend in using the data analysis programs, such as R and WEKA. The use of ML in nutrition is recent and challenging. Therefore, it is encouraged that more studies are carried out relating these themes for the development of food reeducation programs and public policies.

KEYWORDS

Food intake; diet; artificial intelligence; machine learning; supervised and unsupervised algorithms; computational tools

Introduction

Unhealthy food intake is an important behavioral risk factor for many noncommunicable diseases (NCD), including cardiovascular disease (CVD), cancer, and diabetes mellitus (Dao et al. 2019). For this reason, the evaluation of the food intake of an individual or a population plays an important role in the search for answers and a better understanding between the relationship of diet and the triggering of these diseases, intending to make dietary recommendations more effective and promoting public policies (Shim, Oh, and Kim 2014; Rupasinghe, Perera, and Wickramaratne 2020). Several methods of food intake evaluation are available and each has different positive and negative points. The most commonly used are the Food Frequency Questionnaires (FFQ), 24-hour recall (R24H), food record, and food history. However, these methods' imprecision poses a major challenge to the understanding between diet and NCD (Shim, Oh, and Kim 2014).

In this sense, information technologies are being increasingly used in several areas, including health, to support scientific research and clinical decision-making (Ma and Chen 2019). As healthcare institutions generate and store large volumes of data daily, clinical decisions should not be made only based on the intuition and experience of the healthcare professional, but also on the knowledge stored overtime in their databases (Singh, Singh, and Pandi-Jai 2018). Thus, motivated by the need to extract useful information from the collected data, Machine Learning (ML) algorithms have been applied with the potential to generate knowledge that can help improve the quality of service provided to patients, reduce morbidity, mortality, the length of stay and hospital costs, improving the quality of life of patients (Ma and Chen 2019).

ML is a subarea of Artificial Intelligence (AI) whose objective is to detect and extract hidden information, patterns, and specific data that were previously unidentified in large volumes of data because they were difficult to investigate manually or cannot be detected with conventional statistical methods (Assari, Azimi, and Taghva 2017). This knowledge discovery can be automatic or semi-automatic through algorithms that detect and extract information in datasets quickly and accurately (Siqueira-Batista and Silva 2019). ML algorithms can be classified as supervised, unsupervised, semi-supervised, and reinforcement learning approaches and have a wide application in several fields, such as space technology, criminal investigation, bioinformatics, economics, business, among others (Kodati, Vivekanandam, and Ravi 2019). Those algorithms have been applied in the health area to help defining policies of prevention, prediction, diagnosis, early prognosis of diseases, and appropriate and effective treatments (Al-Maqaleh and Abdullah 2017; Babu et al. 2017).

It is noteworthy that using ML algorithms in the health field is still challenging, as health professionals need to

CONTACT Larissa Oliveira Chaves larissaochaves@yahoo.com.br

accurately interpret information derived from the algorithms in a clinical setting and epidemiological studies. In addition, there are still studies to be carried out involving the application of ML algorithms in the area of nutrition, especially in food intake. Therefore, this review can serve as a guide for health professionals interested in food intake and data science, and to enlighten as well as to assist other researchers in the development of new studies. This systematic review aimed to identify and analyze original articles that applied ML algorithms of supervised and unsupervised approaches to assess food consumption in different populations.

Methods

Protocol and registration

This review was conducted in accordance with Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) (Liberati et al. 2009) and was registered in the PROSPERO database (www.crd.york.ac.uk/prospero/), registration number CRD42020198633.

Literature research

Two authors (L.O.C and A.L.G.D) independently searched for original articles that used ML algorithms to evaluate food intake using the following electronic databases: MEDLINE (PubMed, www.pubmed.com), Lilacs (www.lilacs.bvsalud.org), Science Direct (www.sciencedirect.com), SciELO (www.scielo.org) and Google Scholar (https://scholar.google.com.br/). The following descriptors were used as a strategy for research in titles and abstracts: ("machine learning" OR "deep learning" OR "data mining" OR "unsupervised learning") AND ("food intake" OR "diet" OR "food pattern" OR "dietary pattern" OR "food frequency consumption" OR "food questionnaire") NOT review.

The research strategy was not restricted by publication year and language. The research was conducted between July 1st and 6th, 2020. A reverse search was conducted to identify relevant articles cited in the selected studies.

Eligibility criteria

The following criteria were applied for the inclusion of studies: (1) original research; (2) studies in humans; (3) evaluation of food consumption; (4) use of ML algorithms. The following exclusion criteria were applied: (1) non-original publications, such as reviews, letters, book chapters, case reports, abstracts, and comments; (2) animal studies; (3) in vitro studies; (4) research on the composition and importance of a specific food; (5) research that did not use ML algorithms to assess food consumption.

Study selection and data collection process

The studies' selection was made by two authors (L.O.C and A.L.G.D) in three phases: analysis of titles, abstracts, and full texts, independently. After reading the selected studies, a comparison of the compiled data to ensure their integrity and reliability was conducted by the authors. Divergent decisions were resolved by consensus or by consultation with a third author (D.L.F). For each included study, the following information was extracted: authors, year of publication, the country where the research was developed, the objective of the study, characteristics of the participants, method of food intake evaluation, ML approach and algorithms, and computational tools.

Data analysis

All studies selected in this article are summarized in Table 1 according to their main characteristics. The studies were organized chronologically by year of publication, starting with the first published study. The year of publication, location of the study, methods for assessing food consumption, and ML approaches and algorithms along with the computational tools used were considered the main characteristics of this systematic review.

The performance of a meta-analysis was not justified due to the heterogeneity among the studies included. Therefore, according to the Cochrane manual, the authors performed a systematic review (Higgins and Green 2011).

Results

Study selection

A total of 252 studies were identified through the searches in the databases. After the removal of 49 duplicate studies, 203 unique records remained, of which 133 studies were excluded based on their titles and abstracts because they were considered irrelevant: 72 did not use ML algorithms to evaluate food consumption, 45 were animal studies, 11 were not original articles, 3 did not use ML algorithms and 2 were in vitro studies. The remaining 70 studies were reviewed and evaluated in full for eligibility, and 36 met all the criteria adopted for this systematic review and were thus included (Figure 1).

Description of included studies

Overview of the number of publications in relation to the year and countries of studies

The first publication analyzed in this review was in 2008, with only nine publications between 2008 and 2016. From 2017 there was an increase in research in this area, with 27 studies published between 2017 and 2020 (up to the time of this review) as presented in Table 2 and Figure 2.

Of the 36 studies included in this review, 38.9% were conducted in North America, 30.6% in Europe, 25% in Asia, and 2.8% in Oceania. Table 2 shows the distribution of publications in relation to the countries where the studies were conducted. Only one study used data from 12 countries, but they were not mentioned (Yu et al. 2020).

The heatmap in Figure 3 illustrates the relationship of publications from 2008 to 2020 with the countries where the

Table 1. Characteristics of the studies that applied ML algorithms to assess food intake in different populations.

Reference/Country	Objective	Population characteristics	Method for assessing food intake	ML Approach	ML algorithms	Computational tools
Hearty and Gibney (2008) Ireland	Assess ML algorithms to predict HEI based on food intake	1,379 men and women Age: 18 to 64 years	- Registration of daily intake for 7 days, with the times of each meal - HEI divided into quintiles 1 and 5	Supervised	- ANN - Decision Tree	SPSS
De Cos Juez et al. (2009) Spain	Develop mathematical method to predict BMD of women in post-menopause, according to nutritional variables	305 healthy postmenopausal women Age: 50 to 69 years	- Specific questionnaire taken from a food history questionnaire, type FFQ	Supervised	-MARS	Does not contain this information
Ordóñez et al. (2009) Spain	Determine factors that influence the BMD, for the development of specific prevention programs	305 healthy postmenopausal women Age: 49 to 69 years	- DHQ (FFQ type)	Supervised	- SVM - CARTs	Does not contain this information
Zenitani, Nishiuchi, and Kiuchi (2010) Japan	Analyze and extract food patterns	634 employees of a Tokyo electricity company	- Auto Meal Record System (purchase of meals in 2 company cafeterias for 1 year) - Employee ID, date/ time of purchase, quantity and price	Supervised	- Multiple Linear Regression	SAS
De Cos Juez et al. (2011) Spain	Develop BMD prediction method for post-menopausal women, according to nutritional data	200 healthy post-menopausal women Age: 50 to 69 years	- DHQ (FFQ type)	Supervised	- MLP - GA	R
Lazarou et al. (2012) Republic of Cyprus	Applying the ML approach to ascertain eating habits related to childhood obesity	1,140 children Age: 9 to 13 years	- FFQ semi-quantitative - FGFQ: 15 food groups - SEBBQ: 8 psychological aspects of eating - SDHQ: 19 eating habits	Supervised	- Decision Tree	WEKA
Silvera et al. (2014) United States	Examine the interactions between diet, lifestyle, and medical factors with risks of esophageal and gastric cancer subtypes	1,095 cases and 687 controls, men and women Age: 30 to 79 years	- FFQ validated	Supervised	- Decision Tree	CART
Zeevi et al. (2015) Israel	Quantify the PPGR, characterize its variability and identify associated factors	900 healthy men and women Age: 18 to 70 years	- Own questionnaire - Registration on smartphone of meals for 7 days	Supervised	- Stochastic Gradient Boosting Regression	Does not contain this information
Giabbanelli and Adams (2016) United Kingdom	Evaluate food intake by ML and investigate its performance to predict and make food recommendations	4,156 (2,083 adults and 2,073 children) NDNS data 2008-2012	- Registration in NDNS journals	Supervised	- Decision Tree	WEKA
Dipnall et al. (2017) United States	Develop a method to evaluate the risk of depression using GSEM, comparing ML with statistical analysis	5,546 men and women Age: 18 to 80 years. (Data from NHANES 2009-2010)	— 1st interview: R24H - 2nd interview: by phone	Supervised	- GSEM probitmodel - RL	STATA
Kanerva et al. (2018) Finland	Explore sociodemographic and lifestyle risk factors related to overweight using ML and LR	6,258 men and women Age: 25 to 74 years	- FFQ validated and self-managed	Supervised	- RF - LR	R

(continued)

Table 1. Continued.

Reference/Country	Objective	Population characteristics	Method for assessing food intake	ML Approach	ML algorithms	Computational tools
Mezgec and Seljak (2017) Slovenia	Introducing a new approach to food and beverage image detection and recognition	225,953 images, 512 × 512 pixel, from 520 classes of food and beverage images	- Food and beverages from dietary evaluation system of PD Nutrition - Food and beverage image search on Google	Supervised	- NutriNet	Google Custom Search API
Mutter et al. (2017) China	Understand the interaction between nutritional and socioeconomic risk factors with anemia	2,849 men and women Age: 20 to 87 years Five-year follow-up: 1,262 individuals	- Three consecutive days food record, including one day of the weekend	Unsupervised	- SOM	R
Silva et al. (2018) United States/Canada	Introducing new mobile system that enables automated image-based food recognition and evaluation for dietary intervention	Creation of application to evaluate diet through food images and perform interventions	- Image capture of food by smartphone	Supervised	- SVM - Inception V3	Google Colaboratory
Easton, Sicilia, and Stephens (2019) Mexico	Investigate whether specific food groups can predict and classify adults with obesity, diabetes or both	11,385 men and women Age: 20 to 69 years	- FFQ applied by interviewer	Supervised	- Naïve Bayes	Does not contain this information
Forman, Goldstein, Zhang, et al. (2019) United States	Evaluate the feasibility, effectiveness, and acceptability of OT, food lapses and weight loss	43 men and women Age: 18 to 65 years With overweight or obesity	- Registration feeds in the WW - Registration of lapses and triggers in OT for 8 weeks	Supervised	- OT: set of algorithms (Logitboost, Bagging, Random Subspace, RF, and Bayes Network)	Does not contain this information
Guan et al. (2018) Australia	Identify food choices for meals of overweight and obese volunteers for a weight loss test	433 (116 men and 317 women) Average age of 43 years With overweight or obesity	- Food history applied by a nutritionist	Unsupervised	- Apriori - Hierarchical Clustering	R
Jia et al. (2019) United States	Develop an AI-based algorithm to automatically detect food items from images	Free living individuals, application users. No sampling	- "eButton"application	Supervised	- CNN	Does not contain this information
Rosso and Giabbanelli (2018) United Kingdom	Assess whether national surveys can be simplified by recording food intake only	4,156 children and adults of both sexes Age: 18 months and over (NDNS data (2008-2012)	- Evaluated daily by means of home measurements and the weight of food and beverages, for a few days	Supervised	- Decision Tree	WEKA
Shiao et al. (2018a) United States	Examine five folate pathway genes and food parameters related to CRC, measuring HEI in families	106 men and women (53 CRC patients and 53 family members and friends) Age: 18 to 80 years	- FFQ - HEI	Supervised	- LR - GR Elastic Net	SAS
Shiao et al. (2018b) United States	Examine healthy eating predictors, using HEI and GI, in families with CRC	106 men and women (53 CRC patients and 53 family members and friends) Age: 18 to 80 years	- FFQ - HEI	Supervised	- LR - GR Elastic Net	SAS
Shiokawa, Date, and Kikuchi (2018) Japan	Describe a KPCA to extract useful information from metabolic profile	386 NMR datasets and 386 ICP-OES of urine samples collected	— 309 nutritional sets of daily food intake records obtained from studies	Supervised	- RF	R e MATLAB
Panaretos et al. (2018) Greece	Compare statistical and ML analyses in the association of food standards and CVD risk	2.020 men and women Adults and elderly	- FFQ semi-quantitative, validated.	Supervised	- *k*NN - RF	R

			- Questionnaire (EPIC)-Greek			
Faruqui et al. (2019) United States	Predict daily glucose levels in T2D patients based on diet, physical activity, weight and glucose level the previous day	10 adults Overweight or obesity and T2D	- Smartphone application for food intake monitoring	Supervised	- LSTM	Does not contain this information
Forman, Goldstein, Crochiere, et al. (2019) United States	Examine the effectiveness of weight loss, participant satisfaction and the frequency of OT lapses. To verify the precision of interventions by OT algorithm	181 men and women Age: 18 to 70 years Overweight or obesity	- Smartphone applications for 10 weeks	Supervised	OT - Set of algorithms (Logitboost, Bagging, Random Subspace, RF, Bayes Net)	R
Hamad et al. (2019) United States	Identify new and existing SNAP participants and examine differences in socio-demographic, health, nutrition and food purchasing behavior	PSDI – 21,806 men and women. Age: 18 years or older of the waves from 1999 to 2013 Food APS – 4,775 families, including 4,548 children and 9,607 adults (men and women). Held between 2012 and 2013	- FAH - FAFH - HEI	Supervised	- Lasso Regression	R
Shao et al. (2019) China	To investigate the effects of smoking, diet and physical activity on the risk of CVD and to propose a new model of risk analysis	23,682 (12,605 men and 11,077 women) Over 50 years of age	- Application to investigate the consumption of alcohol, meat, milk, vegetables and fruit	Supervised	- Decision Tree - RF	WEKA
Yu et al. (2020) 12 countries	Demonstrate the use of ML methods to find food groups related to the CBB	31,551 men and women 8,320 Central Bank cases and 23,231 non-CB cases	- FFQ validated	Supervised	- Decision Tree	R
Burgermaster et al. (2020) United States	Investigate applicability of an application for dietary recommendations in individuals with T2D compared with the specialist	4 patients with T2D	- Smartphone application (photos of meals with description and ingredients for 30 days)	Supervised	- Decision Tree	R
He et al. (2020) United Kingdom	Analyze the causal relationship between lifestyle factors, health indicators and diet	100,000 men and women Over 18 years of age	- Average quantity in kg of cereals, fruits, vegetables and cheese consumption, per year	Supervised	- GTM - GPLVM	MATLAB
Kwon et al. (2020) South Korea	Identify subgroups in the population based on nutritional factors and risk factors for ALM decrease using ML	10,863 men and women Over 40 years old	- R24H applied by nutrition	Supervised and Unsupervised	- *K*-means - Multivariate Logistic Regression	R
Jiang et al. (2020) Denmark	Identify associations between diets and anthropometry	23,195 women and 20,595 men Age: 50 to 64 years	- FFQ	Unsupervised	- Compass-2	Does not contain this information
Xu et al. (2020) United States	Assess the risks of metabolic syndrome in adults with depression, treated with antipsychotics and examine the ability of DQI to predict cardiometabolic risks	106 men and women non-diabetic with depression and 106 men and women non-depressed controls. Age: 18 to 25 years	- Own questionnaire - Hedonic facial scales and horizontal lines: research of likes and dislikes, level of likes and dislikes and level of satisfaction - DQI	Supervised	- Multivariate Regressions	STATA

(*continued*)

Table 1. Continued.

Reference/Country	Objective	Population characteristics	Method for assessing food intake	ML Approach	ML algorithms	Computational tools
Bodnar et al. (2020) United States	Estimate associations between fruit and vegetable intake in relation to adverse pregnancy outcomes	7,572 women	- FFQ	Supervised	- Super Learner com TMLE	Does not contain this information
Narziev et al. (2020) South Korea	Build ML model for depression detection and classification using smartphone	21 students: no depression (n = 5), light (n = 6), moderate (n = 6) and severe (n = 4). For 4 weeks	- STDD smartphone application	Supervised	- SVM - RF	WEKA
Iwendi et al. (2020) China	Recommending diets using deep learning in medical data, to detect which food should be administered	30 men and women of hospitals	- Food intake data available from the hospital database	Supervised	- MLP, RNN, GRU, LSTM, LR, Naïve Bayes RF	Google Colaboratory

Abbreviations: AI: Artificial Intelligence; ALM: Appendicular Lean Mass; ANN: Artificial Neural Networks; BC: Bladder Cancer; BMD: Bone Mineral Density; CARTs: Classification and Regression Trees; CNN: Convolutional Neural Network; CRC: Predictors of Colorectal Cancer; CVD: Cardiovascular Disease; DHQ: Diet History Questionnaire; DQI: Diet Quality Indexes; EPIC: European Prospective Investigation into Cancer and Nutrition; FAH: Food at Home; FAFH: Food Away From Home; FFQ: Food Frequency Questionnaire; FGFQ: Food Groups Frequency Questionnaire; GA: Genetic Algorithms; GI: Glycemic Index; GPLVM: Gaussian Process Latent Variable Model; GR: Generalized Regression; GRU: Gated Recurrent Units; GSEM: Generalized Structural Equation Model; GTM: Generative Topological Mapping; HEI: Healthy Eating Index; ICP-OES: Inductively Coupled Plasma Optical Emission Spectrometry; *k*-NN: *k*-Nearest Neighbors; KPCA: Kernel Principal Component Analysis; LR: Logistic Regression; LSTM: Long Short-Term Memory; MARS: Multivariate Adaptive Regression Splines; ML: Machine Learning; MLP: Multilayer Perceptron; NDNS: National Diet and Nutrition Survey; NHANES: National Health and Nutrition Examination Survey; NMR: Nuclear Magnetic Resonance; OT: On Track; PCA: Principal Component Analysis; PPGR: Postprandial Glycemic Response; PSID: Panel Study of Income Dynamics; RF: Random Forest; RNN: Recurrent Neural Network; R24H: 24-h recall; SDHQ: Short Dietary Habits Questionnaire; SEBBQ: Short Eating Habits Behaviors & Beliefs Questionnaire; SNAP: Supplemental Nutrition Assistance Program; SOM: Self-Organizing Map; STDD: Short-Term Depression Detector; SVM: Support Vector Machine; T2D: Type 2 Diabetes; TMLE: Targeted Maximum Likelihood Estimation; WW: Weight Watchers.

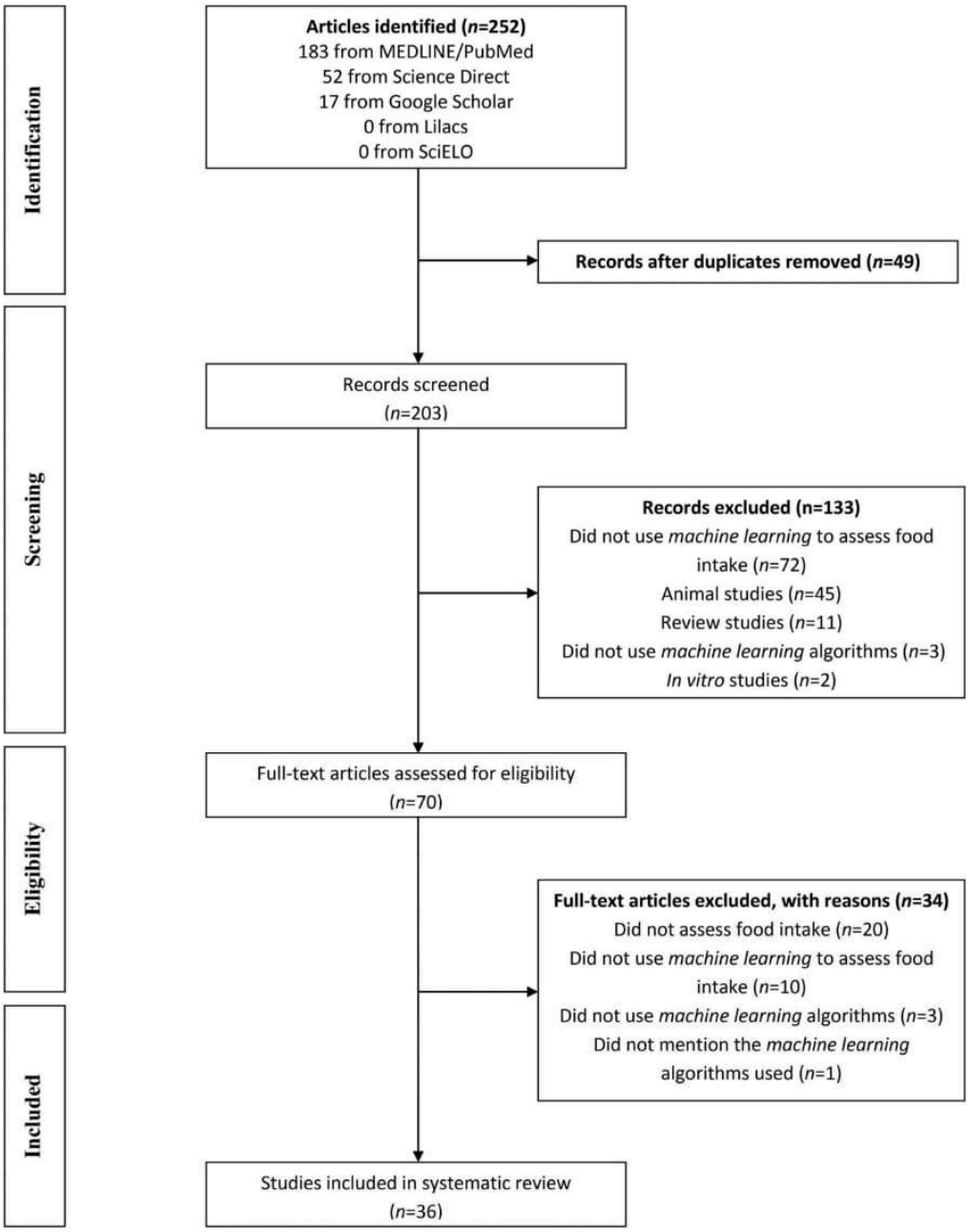


Figure 1. Flowchart of the study selection process, according to PRISMA recommendation.

studies were conducted. For a better visual analysis of these results, the webColorBrewer 2.0 tool was used to choose the color palette that could also be differentiated by colorblind individuals.

Population characteristics

As shown in Table 2, five studies (13.8%) did not describe the population characteristics for gender and age. Four studies (11.1%) included only women, and 25 studies (69.4%) included individuals of both sexes. Regarding the age group, 20 studies(55.5%) were conducted with adult and elderly population, three studies (8.3%) with adults, two studies (5.6%) with children and adults, and only one study (2.8%) with children (Lazarou et al. 2012). It was observed that the studies included mostly individuals with overweight or obesity (11.1%) and with a diagnosis of cancer (11.1%), followed by studies with postmenopausal women (8.3%) and

Table 2. Characteristics of the population and publications in relation to the year and countries in which the studies were conducted.

Characteristics	References
Publications in relation to the year	
First: 2008	Hearty and Gibney 2008
2008–2016	Hearty and Gibney 2008; De Cos Juez et al. 2009; Ordóñez et al. 2009; Zenitani, Nishiuchi, and Kiuchi 2010; De Cos Juez et al. 2011; Lazarou et al. 2012; Silvera et al. 2014; Zeevi et al. 2015; Giabbanelli and Adams 2016
2017–2020	Dipnall et al. 2017; Kanerva et al. 2018; Mezgec and Seljak 2017; Mutter et al. 2017; Silva et al. 2018; Easton, Sicilia, and Stephens 2019; Forman, Goldstein, Zhang, et al. 2019; Guan et al. 2018; Jia et al. 2019; Rosso and Giabbanelli 2018; Shiao et al. 2018a; Shiao et al. 2018b; Shiokawa, Date, and Kikuchi 2018; Panaretos et al. 2018; Faruqui et al. 2019; Forman, Goldstein, Crochiere, et al. 2019; Hamad et al. 2019; Shao et al. 2019; Yu et al. 2020; Burgermaster et al. 2020; He et al. 2020; Kwon et al. 2020; Jiang et al. 2020; Xu et al. 2020; Bodnar et al. 2020; Narziev et al. 2020; Iwendi et al. 2020
Publications in relation to the countries	
Australia	Guan et al. 2018
China	Mutter et al. 2017; Shao et al. 2019; Iwendi et al. 2020
Denmark	Jiang et al. 2020
Finland	Kanerva et al. 2018
Greece	Panaretos et al. 2018
Ireland	Hearty and Gibney 2008
Israel	Zeevi et al. 2015
Japan	Zenitani, Nishiuchi, and Kiuchi 2010; Shiokawa, Date, and Kikuchi 2018
Mexico	Easton, Sicilia, and Stephens 2019
Republic of Cyprus	Lazarou et al. 2012
Slovenia	Mezgec and Seljak 2017
South Korea	Kwon et al. 2020; Narziev et al. 2020
Spain	De Cos Juez et al. 2009; Ordóñez et al. 2009; De Cos Juez et al. 2011
United Kingdom	Giabbanelli and Adams 2016; Rosso and Giabbanelli 2018; He et al. 2020
United States	Silvera et al. 2014; Dipnall et al. 2017; Forman, Goldstein, Zhang, et al. 2019; Shiao et al. 2018a; Shiao et al. 2018b; Faruqui et al. 2019; Forman, Goldstein, Crochiere, et al. 2019; Hamad et al. 2019; Burgermaster et al. 2020; Xu et al. 2020; Bodnar et al. 2020
United States and Canada	Silva et al. 2018
Population characteristics	
Adults	Faruqui et al. 2019; Xu et al. 2020; Narziev et al. 2020
Adult and elderly	Hearty and Gibney 2008; De Cos Juez et al. 2009; Ordóñez et al. 2009; De Cos Juez et al. 2011; Silvera et al. 2014; Zeevi et al. 2015; Dipnall et al. 2017; Kanerva et al. 2018; Mutter et al. 2017; Easton, Sicilia, and Stephens 2019; Forman, Goldstein, Zhang, et al. 2019; Guan et al. 2018; Shiao et al. 2018a; Shiao et al. 2018b; Panaretos et al. 2018; Forman, Goldstein, Crochiere, et al. 2019; Shao et al. 2019; He et al. 2020; Kwon et al. 2020; Jiang et al. 2020
Anemia	Mutter et al. 2017
Both sexes	Hearty and Gibney 2008; Zenitani, Nishiuchi, and Kiuchi 2010; Lazarou et al. 2012; Silvera et al. 2014; Zeevi et al. 2015; Giabbanelli and Adams 2016; Dipnall et al. 2017; Kanerva et al. 2018 ; Mutter et al. 2017; Easton, Sicilia, and Stephens 2019; Forman, Goldstein, Zhang, et al. 2019; Guan et al. 2018; Rosso and Giabbanelli 2018; Shiao et al. 2018a; Shiao et al. 2018b; Panaretos et al. 2018 ; Faruqui et al. 2019; Forman, Goldstein, Crochiere, et al. 2019; Shao et al. 2019; Yu et al. 2020; He et al. 2020; Kwon et al. 2020; Jiang et al. 2020; Xu et al. 2020; Narziev et al. 2020
Cancer	Silvera et al. 2014; Shiao et al. 2018a; Shiao et al. 2018b; Yu et al. 2020
Children	Lazarou et al. 2012
Children and adults	Giabbanelli and Adams 2016; Rosso and Giabbanelli 2018
Depression	Xu et al. 2020; Bodnar et al. 2020
Did not describe	Jia et al. 2019; Shiokawa, Date, and Kikuchi 2018; Hamad et al. 2019; Iwendi et al. 2020; Burgermaster et al. 2020
Overweight and / or obesity	Kanerva et al. 2018; Forman, Goldstein, Zhang, et al. 2019; Guan et al. 2018; Forman, Goldstein, Crochiere, et al. 2019
Postmenopausal women	De Cos Juez et al. 2009; Ordóñez et al. 2009; De Cos Juez et al. 2011
Type 2 diabetes mellitus	Burgermaster et al. 2020
Two or more noncommunicable diseases	Easton, Sicilia, and Stephens 2019; Faruqui et al. 2019
Women	De Cos Juez et al. 2009; Ordóñez et al. 2009; De Cos Juez et al. 2011; Bodnar et al. 2020

individuals diagnosed with depression (5.6%), anemia (2.8%) and type 2 diabetes mellitus (2.8%). Only two studies (5.6%) investigated a population with two or more NCD. The other studies did not inform the population characteristics. Note that most of the studies included in this review were with a population composed of both sexes, adults and elderly, with the presence of overweight, obesity, or cancer.

Methods of food intake evaluation

A total of 13 studies (36.1%) used the FFQ as a method to evaluate food consumption, followed by ten studies (27.8%) that used smartphone/software applications. Five studies (13.9%) used other types of questionnaires, four studies (11.1%) used the food registry, two studies (5.6%) used the R24H applied by a trained interviewer, and only one study (2.8%) used hedonic scales. Four studies (11.1%) of the selected ones used food intake data recorded in database systems and/or studies already published. One study (2.8%) aimed at detecting and recognizing food and beverage images used Google site as a search tool (Table 3).

ML algorithms and computational tools

The studies included in this review used different ML algorithms, and the most used ones were in the category supervised learning. Of the 36 studies included, 32 studies (88.9%) used supervised approach algorithms, being 23 studies (63.9%) of the classification type and 14 studies (38.9%) of the regression type. The most used classification algorithms were based on Decision Trees with 13 studies (36.1%), and the Artificial Neural Networks with 6 studies

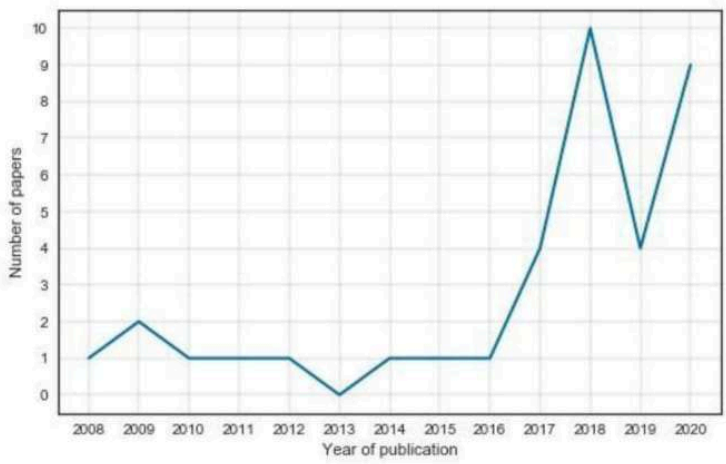


Figure 2. Number of publications per year.

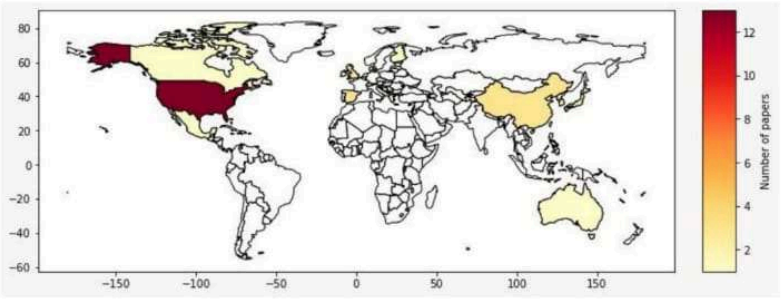


Figure 3. Heatmap relating the number of articles published in the countries where the studies were conducted.

(16.7%). We found only four studies (11.1%) in which unsupervised approaches were applied. The clustering algorithms found were Hierarchical clustering, *k*-means, and Self-Organizing Maps, while the Apriori algorithm was the only association rule procedure encountered. Only one study (2.8%) applied both ML approaches, supervised and unsupervised (Kwon et al. 2020) (Table 3).

Nine studies (25%) did not inform the computational tools used. However, ten studies (27.8%) used the Statistical Program R, five studies (13.9%) used the WEKA program, three studies (8.3%) used SAS, two studies (5.6%) used STATA, two studies (5.6%) used Google Colaboratory, one study (2.8%) used CART, one study (2.8%) MATLAB, one study (2.8%) SPSS and one study (2.8%) Google Custom Search API. Only one study (2.8%) used two computational tools, R and MATLAB (Table 3).

Discussion

Overview of growth in studies and publications involving ML and nutrition

According to our knowledge, this was the first systematic review that analyzed the application of different ML algorithms, supervised and unsupervised approaches, to evaluate food intake in healthy and unhealthy individuals. Currently, there is considerable scientific interest in the use of those algorithms due to the high predictive performance in large volumes of data, such as in agriculture, transport, finance, criminal justice, and health (Cutillo et al. 2020). In the health area, ML algorithms have great potential to improve the results of patients from clinical research to hospital care, helping in the process of diagnosis and prediction of diseases (Cutillo et al. 2020; Fernandes and Filho 2019).

This growth in studies and publications involving the application of ML algorithms in health is confirmed in our review due to the significant increase in studies that addressed ML and food intake from 2016 onwards with the start of the peak in 2017. Of the 36 studies included, 27 were published between 2017 and mid-2020. Moreover, we observed an increase since 2011 in studies related to AI in the health area, with peaks also from the year 2017 onwards (data not shown). This observation emphasizes that the use of computational methods is not restricted only to nutrition. Therefore, there is a trend in the use of AI in the health area in general.

Even though there has been a recent increase in the number of publications addressing the various applications of ML algorithms in nutrition, the use of artificial intelligence approaches in other areas has been under discussion for

Table 3. Characteristics of included studies in relation to the method of assessing food intake and the type of algorithm and computational tools.

Assessment methods of food intake	
Food consumption data recorded in database systems and / or studies already published	Silvera et al. 2014; Giabbanelli and Adams 2016; Shiokawa, Date, and Kikuchi 2018; Iwendi et al. 2020
Food Frequency Questionnaires	De Cos Juez et al. 2009; Ordóñez et al. 2009; De Cos Juez et al. 2011; Lazarou et al. 2012; Silvera et al. 2014; Kanerva et al. 2018; Easton, Sicilia, and Stephens 2019; Shiao et al. 2018a; Shiao et al. 2018b; Panaretos et al. 2018; Yu et al. 2020; Jiang et al. 2020; Bodnar et al. 2020
Food registry	Hearty and Gibney 2008; Mutter et al. 2017; Rosso and Giabbanelli 2018; He et al. 2020
Google as a search tool	Mezgec and Seljak 2017
Hedonic scales	Xu et al. 2020
24-hour recall	Dipnall et al. 2017; Kwon et al. 2020
Other types of questionnaires	Lazarou et al. 2012; Zeevi et al. 2015; Guan et al. 2018; Hamad et al. 2019; Xu et al. 2020
Smartphone / software applications	Zenitani, Nishiuchi, and Kiuchi 2010; Zeevi et al. 2015; Silva et al. 2018; Forman, Goldstein, Zhang, et al. 2019; Jia et al. 2019; Faruqui et al. 2019; Forman, Goldstein, Crochiere, et al. 2019; Shao et al. 2019; Burgermaster et al. 2020; Narziev et al. 2020
Supervised approach algorithms	
Type: classification	Hearty and Gibney 2008; Ordóñez et al. 2009; De Cos Juez et al. 2011; Lazarou et al. 2012; Silvera et al. 2014; Giabbanelli and Adams 2016; Kanerva et al. 2018; Mezgec and Seljak 2017; Silva et al. 2018; Easton, Sicilia, and Stephens 2019; Forman, Goldstein, Zhang, et al. 2019; Jia et al. 2019; Rosso and Giabbanelli 2018; Shiokawa, Date, and Kikuchi 2018; Panaretos et al. 2018; Faruqui et al. 2019; Forman, Goldstein, Crochiere, et al. 2019; Shao et al. 2019; Yu et al. 2020; Burgermaster et al. 2020; Bodnar et al. 2020; Narziev et al. 2020; Iwendi et al. 2020
Type: regression	De Cos Juez et al. 2009; Zenitani, Nishiuchi, and Kiuchi 2010; Zeevi et al. 2015; Dipnall et al. 2017; Kanerva et al. 2018; Forman, Goldstein, Zhang, et al. 2019; Shiao et al. 2018a; Shiao et al. 2018b; Forman, Goldstein, Crochiere, et al. 2019; Hamad et al. 2019; He et al. 2020; Kwon et al. 2020; Xu et al. 2020; Iwendi et al. 2020
Artificial Neural Networks	Hearty and Gibney 2008; De Cos Juez et al. 2011; Mezgec and Seljak 2017; Silva et al. 2018; Jia et al. 2019; Faruqui et al. 2019
Decision Trees	Hearty and Gibney 2008; Ordóñez et al. 2009; Lazarou et al. 2012; Silvera et al. 2014; Giabbanelli and Adams2016; Kanerva et al. 2018; Rosso and Giabbanelli 2018; Shiokawa, Date, and Kikuchi 2018; Panaretos et al. 2018; Shao et al. 2019; Yu et al. 2020; Burgermaster et al. 2020; Narziev et al. 2020
Unsupervised approach algorithms	
Apriori	Guan et al. 2018; Jiang et al. 2020
Hierarchical Clustering	Guan et al. 2018
K-means	Kwon et al. 2020
Self-Organizing Map	Mutter et al. 2017; Jiang et al. 2020
Computational tools	
CART	Silvera et al. 2014
Did not inform	De Cos Juez et al. 2009; Ordóñez et al. 2009; Zeevi et al. 2015; Easton, Sicilia, and Stephens 2019; Forman, Goldstein, Zhang, et al. 2019; Jia et al. 2019; Faruqui et al. 2019; Jiang et al. 2020; Bodnar et al. 2020
Google Colaboratory	Iwendi et al. 2020; Silva et al. 2018
Google Custom Search API	Mezgec and Seljak 2017
MATLAB	He et al. 2020
R	De Cos Juez et al. 2011; Kanerva et al. 2018; Mutter et al. 2017; Guan et al. 2018; Panaretos et al. 2018; Forman, Goldstein, Crochiere, et al. 2019; Hamad et al. 2019; Yu et al. 2020; Burgermaster et al. 2020; Kwon et al. 2020
SAS	Zenitani, Nishiuchi, and Kiuchi 2010; Shiao et al. 2018a; Shiao et al. 2018b
SPSS	Hearty and Gibney 2008
STATA	Dipnall et al. 2017; Xu et al. 2020
Statistical Program R and MATLAB	Shiokawa, Date, and Kikuchi 2018
WEKA	Lazarou et al. 2012; Giabbanelli and Adams 2016; Rosso and Giabbanelli 2018; Shao et al. 2019; Narziev et al. 2020

many years (Smallwood and Sondik 1973). One of the possible explanations for the increase in the use of ML algorithms in nutrition and health in general in recent years is precisely the search for more accurate procedures to meet the needs of professionals in their daily decision-making activities, treatment options, and reduction of health costs (Reis et al. 2017). In the long term, it is believed that ML approaches will benefit professionals in diverse fields, by offering objective suggestions and ways to improve the efficiency, reliability, and accuracy of processes.

Influence of regionalization on food consumption

According to the results achieved, a small diversity of countries investigating food intake using ML algorithms was noted. Most studies were conducted in North America, followed by Europe, Asia, and Oceania, and no studies were developed in Central America, South America, and Africa. It is known that food intake is highly influenced by the region in which we live, so it is important that countries can conduct their research to understand the eating behavior of individuals in the same region (Latha and Thegaleesan 2019).

The food intake pattern is directly influenced by social, cultural, and economic factors (Savage, Bambrick, and Gallegos 2020). In addition, the characteristics of the population in terms of customs, level of education, knowledge about healthy eating, workplace, family and friends circle also have a major impact on food choice and habits (Latha and Thegaleesan 2019). These differences can be observed between different countries or regions within the same country (Vasileska and Rechkoska 2012).

In this review, an interesting result found is that the United States was the country that developed the largest number of studies involving ML and food consumption. It is believed that this great interest in research on nutrition is related to the low quality of the diet consumed and also to the reduction of physical activity practices of its inhabitants, which has been worsening since the 1980s (Popkin, Adair, and Ng 2012). Economic development and increasing urbanization in developed countries, such as the United States, brought benefits and negative consequences for lifestyle and dietary patterns, which include quantitative and qualitative changes in the diet. This more industrialized dietary pattern includes an increase in the consumption of high-calorie foods, refined carbohydrates, and saturated fats of animal origin, in addition to a reduction in the intake of complex carbohydrates, fibers, vitamins, and minerals (Vasileska and Rechkoska 2012).

It is essential to point out that cultural and behavioral factors are also susceptible to change and that the circle of family and friends is extremely important in the correct choice of food. In addition, an increasing number of individuals have been eating outside their homes, which further increases the consumption of processed and ultra-processed foods since access to healthy options is often limited in many places, including at work and in school environments (Latha and Thegaleesan 2019).

Methods of food intake evaluation

The evaluation of adequate and reliable food intake in scientific research is important to understand the association between diet and the health conditions of an individual or a population. It has also been useful in predicting NCD (Vucic et al. 2009). However, an accurate assessment of food intake remains a major challenge, as it is subject to bias, and none of it is considered the gold standard. The most commonly used methods to assess food intake are food histories, food records, R24H, and FFQ (Vucic et al. 2009; Shim, Oh, and Kim 2014).

In this review, most of the selected studies used the FFQ to evaluate food consumption. This method is considered one of the simplest, cheapest, fastest, and easiest to administer and process, and allows for long-term food evaluation (Chmurzynska et al. 2018). This method contains a defined list of about 100 to 150 food items and options of the usual frequency of consumption over the period consulted. In some cases, portion sizes are also investigated, but little information is collected on the additional characteristics of the food consumed (Shim, Oh, and Kim 2014). Despite this methodological limitation, FFQ has been widely used in epidemiological studies since the 1990s. It is important to note that the FFQ should be developed specifically for each study and research group because diet can be influenced by ethnicity, culture, economic status, among others (Shim, Oh, and Kim 2014; Rupasinghe, Perera, and Wickramaratne 2020).

The R24H and food registration have some important limitations that can influence their choices, such as collecting information for a specific period, usually for short-term intake. Thus, to measure the average intake, several R24H or food records are necessary. Repeated measurement requires resources and time and can influence respondents' food intake, improving the quality of the diet, changing or omitting information intentionally (Rupasinghe, Perera, and Wickramaratne 2020). The R24H is conducted by interview and usually requires 20 to 30 minutes, and the information depends on the interviewees' memory and the interviewer's skills. On the other hand, food recording is a method that takes more time to obtain accurate data and respondents must undergo prior training. Therefore, a high level of motivation becomes necessary. Also, each questionnaire requires a thorough review to ensure that all reported data are correct (Shim, Oh, and Kim 2014; Rupasinghe, Perera, and Wickramaratne 2020).

However, both methods to evaluate food intake also have common strengths, such as being easy to apply, having a wide variety of foods, are made up of open questions that allow the collection of great information on consumption and can be used to estimate the average consumption of a given population. Moreover, food registration does not depend on the individual's memory since the information is self-reported when the food is consumed (Shim, Oh, and Kim 2014; Chmurzynska et al. 2018).

As seen above, the most used methods nowadays have many limitations, including memory dependency, understanding of food portions, literacy, and training of interviewers. Motivated by the development of reliable evaluation methods, the technology emerges as a viable solution to current methodological deficiencies with the potential to improve adherence, communication, and data quality (Sharp and Allman-Farinelli 2014). As a result, a large number of studies that used mobile applications were found in our review. In recent years, mobile devices have been used to evaluate individual and group diets in real-time, incorporating their daily food routines. The easy access and interactive features of these applications, such as setting goals and dietary lapses, allow users to monitor the diet and trigger healthier behaviors. Mobile applications have demonstrated validity and reliability, similar to conventional methods. In addition, the use of the application feeds continuous progress data to be used in future studies (Chmurzynska et al. 2018; Ahn et al. 2019).

ML algorithms and computational tools

ML is a subarea of AI whose objective is to develop algorithms that give computers or computer systems the ability to learn specific knowledge, behavior, or pattern automatically or semi-automatically from examples or informed observations (Michalski, Carbonell, and Mitchell 2013). ML approaches can be of types: supervised, unsupervised, semi-supervised and reinforcement learning. Here, we will discuss the first two main approaches, which were the ones found in the studies selected in this review.

Supervised learning approach

In situations where supervised learning is applied, one has prior knowledge of the values of the output variable, i.e., the

classes or labels represented by categorical or continuous values of the input dataset used - composed of registers (instances) and variables (attributes). Therefore, the objective of supervised learning is to learn, employing algorithms for this type of task, a mapping function that best approximates the relationship between input data and observable output so that when new instances are available, the output can be predicted with considerable accuracy (Pedregosa et al. 2011).

This learning process works in the following way: first, the dataset is split into two parts: training and test data. A predictive model is then built based on an algorithm that uses the training set so that the resulting model learns patterns by associating the input data values with the output labels. After the training, the model will receive the test set split, which was left out of the previous step, and it will apply the knowledge learned from previous experiences (training data) to this test set so that the accuracy, sensitivity, specificity – and other important statistical measures – are calculated to evaluate the predictive power of the model (Dey 2016). Thus, together with performance metrics, the model ability to generalize to predict labels for previously unseen instances during the training will be evaluated. (Michalski, Carbonell, and Mitchell 2013).

Interestingly, it was observed that most of the studies included in this review, totaling 32 out of 36, used some supervised learning algorithm. Below, we will discuss these algorithms, which are of the classification and regression type, and the reasons why they were the most used in the studies included in this review.

Classification algorithms. The classification algorithms are used when the goal is to map the input variables to a specific categorical class. It is common to find in the literature works that applies some of the supervised algorithms based on Decision Trees (DT), Artificial Neural Networks (ANN), Naïve Bayes (NB), Support Vector Machine (SVM), Logistic Regression, and *k*-Nearest Neighbor (*k*NN) (Khan et al. 2010). In this review, we provide further details of the algorithms based on DT and ANN, as they are the most present approaches in the selected studies.

DT or derivatives from this approach are quite popular classification algorithms used to build predictive models (Rajput et al. 2011). This type of technique expresses the possible results of a series of choices related to attributes and classes through rules. Each tree is represented through a structure with nodes and branches, and each non-leaf node in the tree is a decision rule. A DT usually uses the top-down approach, i.e., from the root node to leaves (Rajput et al. 2011). It is started with a single root (parent node), representing the most important attribute in the dataset according to an impurity metric. Between the root and the (child nodes), which represent other attributes, the branches or edges connect these nodes and represent the possible values of the attribute analyzed by the predecessor or parent node. Finally, after traversing the tree, one reaches the leaf nodes representing the target, i.e., predicted classes (Dey 2016).

It is believed that the vast demand for this type of algorithm in the review studies is due to its advantages, especially: fast construction of the predictive model; fast classification of new instances; no need for normalization or standardization in the preprocessing phase; simplicity in understanding and interpreting the rules generated even for non-specialist users, as the resulting tree provides a consolidated view of the classification logic (Khan et al. 2010; Rajput et al. 2011).

According to Yu et al. (2020) the application of the DT algorithm in their study strongly contributed to the high accuracy found in the proposed classification, indicating that the ML can adequately deal with missing data and measurements of complex investigation. The investigators concluded that the DT algorithm provided an effective approach to identify some food groups related to bladder cancer risk.

Another very powerful and frequently applied supervised ML approach is ANN. ANN algorithms are inspired by the operating structure of the biological neural system concerning the ability to learn from data and improve its performance according to what was learned through operations such as parallel calculations for data processing and knowledge representation (Tan, Steinbach, and Kumar 2006).

An ANN is built by a set of processing units, also known as neurons, linked by weighted connections or synaptic weights responsible for the propagation of attribute values between the neurons in the layers (Tan, Steinbach, and Kumar 2006; Michalski, Carbonell, and Mitchell 2013). A neuron is a component that calculates the weighted sum of the values received as input, applies an activation function, and passes the result forward to the next layer. The intermediate layers, if any, between the input and output layers, are known as hidden layers. The value propagation process continues until reaching the output layer with the predicted response (Tan, Steinbach, and Kumar 2006; Michalski, Carbonell, and Mitchell 2013).

The main advantage of implementing ANN-based algorithms is the high capacity to learn from large volumes of data, whether structured or not and in diverse applications (e.g., speech recognition, machine translation, image captioning generator, among many others). However, ANN has some disadvantages, such as its high computational cost and physical memory use. Moreover, their training is relatively slow, and the results learned are difficult for users to interpret (Khan et al. 2010).

Many of the classification algorithms, including ANN, are known to be difficult to understand and explained in simple terms how the predictions were made. When built, these models are called black-box models. In the health area, this kind of model is even more challenging because the professionals will be apprehensive in making decisions, especially those related to death risk, without a firm understanding of how the algorithm came to that predicted recommendation (Khan et al. 2010).

A study by Silva et al. (2018) trained an effective food classification model in a food image dataset and found that neural networks achieved an overall performance of 87.2% (with 90.0% sensitivity and specificity of 84.4%). When the

model was trained based on this food image dataset, it achieved a precision of 65.5% (with a sensitivity of 59.0% and a specificity of 72.0%). They concluded that the main contribution of neural networks is that they automatically learn resources through convolutional layers, with high performance and accuracy.

In this context, we strongly believe that the frequent and broad use of DT in the studies addressed in this systematic review was for the speed of training and mainly for providing a clear explanation for the results found by the model (Lundberg and Lee 2017).

Regression algorithms. Regression algorithm sare used in situations where the aim is to map the input variables to an output with a continuous value, i.e., any numerical value between two limits (Kan et al. 2019). Note that regression algorithms as well as classification algorithms were widely used in the studies selected in this review.

The objective of regression is to define the parameter values of a mathematical equation that defines y (the output to be predicted) as a function of variables x (input variables) so that the error concerning the adjusted curve and all the data points is minimized. This equation, the final model, can then be used to predict the result for new instances. In general, a model fits the data well if the differences between the observed values and the predicted values are small and unbiased (Pedregosa et al. 2011).

Among the many forms of regression described in the literature (Multiple Linear Regression, Lasso Regression, among others), one must select the best technique that explains the data to be analyzed. The best way to verify this is by applying different regression models and comparing the performance in predicting for new instances (Goldstein, Navar, and Carter 2017). For regression, we use as a measure of performance the calculated error in relation to the model obtained (curve) and the points belonging to the training set. Thus, the smaller the prediction error, the better the final performance of the model (Goldstein, Navar, and Carter 2017).

The study by Pagamunici et al. (2014) aimed to develop a high nutritional value gluten-free granola and evaluate it during storage using ML techniques such as multivariate analysis and simple linear regression. Over the storage period analyzed, a positive correlation was observed between appearance and general acceptance and the product remained stable in relation to these parameters. The results of this study demonstrate the high contribution and effectiveness of the application of regression analysis. These analyzes enabled a presentation of an innovative predictive report that gives greater prediction accuracy and a better-tuned model to identify significant predictors.

It is important to mention that some of the algorithms used in classification problems, such as DT, Random Forest, SVM and ANN, also work as regression algorithms. However, those algorithms are modified to adapt the desired output type, in this case, a numeric value, not a categorical label (Rodriguez-Galiano et al. 2015).

Unsupervised learning approach

Unlike supervised approaches, the algorithms of unsupervised learning are used to explore unlabeled data, i.e., when instances have no associated value or category. As a result the unsupervised learning algorithms do not aim to make predictions but, instead, to find potentially useful hidden structures and patterns that humans can interpret and that allow a better description and understanding of the data (Tan, Steinbach, and Kumar 2006).

In this approach, the task of the ML algorithms is not to find the right output from the input data but to explore the data and be able to find clusters or make inferences according to the similarities, patterns, and differences found evaluating the attributes of the instances, without any previous training (Tan, Steinbach, and Kumar 2006). The motivation to use this approach is due to its ability to provide initial insights that can then be used for testing scientific hypotheses and conduct research from a starting point for analysis.

Unsupervised learning tasks are typically to find underlying groups (clusters) in the data and/or reveal important associations rules (Dey 2016). In our review, only four studies applied unsupervised ML approaches of which three used clustering algorithms and one used association rules algorithms.

Clustering algorithms. The most common task in unsupervised learning is clustering. In this case, the unlabeled data are analyzed and organized in clusters by their similarities or dissimilarities (Tan, Steinbach, and Kumar 2006). The measurement of how similar or dissimilar the instances are to each other is done using a proximity calculation, such as the Euclidean distance (Pedregosa et al. 2011). The goal is to create a clustering (a set of clusters) where instances in the same cluster are very similar to each other (each cluster is cohesive), while instances in distinct clusters are highly dissimilar (i.e., clusters are well-separated from each other) (Zheng et al. 2019). In a sense, clustering algorithms reveal hidden categories, i.e., each cluster can be thought as a class of its instances (Ghorbani and Ghousi 2019).

The *k*-means algorithm is the most largely used clustering algorithm. To apply this procedure, it is necessary to give as input to the algorithm the number k of clusters sought. Initially, k centroids (center points) are randomly defined. Then, for each following iteration, every instance is associated to its closest centroid, and each centroid is redefined according to the grouped instances (typically, the new centroid location will be the mean point in the cluster). The redefinition of centroid and resulting association of instances continue throughout multiple iterations until the centroids do not change anymore (Jain 2010).

According to the study by Kwon et al. (2020), the application of the k-means algorithm was fundamental to extract important and hidden information, such as the relashionship between total energy and protein intake, which were difficult to distinguish with conventional analyses. The k-means interestingly contributed to the proper formation of clusters and the comparison of risk factors between them. Cluster-specific risk factors were found to include high consumption

of fat and smoking in the men's cluster and low consumption of carbohydrates, protein, fat, and alcohol consumption in the women's cluster.

Self-Organizing Map (SOM) is another well-known clustering technique that was found in the studies selected in this review. SOM is a particular type of unsupervised neural network, where neurons are arranged in a 2-dimensional grid. Throughout the iterations, the neurons gradually agglutinate around regions presenting high density of data points. Therefore, regions with many neurons can be interpreted as clusters. (Fernandes and Filho 2019).

The study by Mutter et al. (2017) used the SOM algorithm to highlight the inherent natural heterogeneity of nutritional profiles and how they are associated with incident anemia in a population setting, showing how nutritional and economic differences between northern and southern Jiangsu predict differences in incident anemia. The authors highlighted the excellent contribution of this algorithm for the complete separation between training and evaluation data, being one of the strengths of the SOM approach, as its architecture avoids overfitting. Transparency is another strong point of the SOM approach, where the process of defining subgroups and investigating their profiles is guided by the user and open to constructive criticism from other observers.

Another clustering technique found in the studies included in this review was Hierarchical clustering, where data are partitioned successively, producing a hierarchical representation of the group. Hierarchical methods require a matrix containing metrics of distance between clusters, this matrix is known as a matrix of similarities between groups. Distance methods between groups are used to calculate proximity values between groups, such as the Euclidean distance. Through the analysis of the dendogram (diagram that shows the hierarchy and the relationship of the clusters in a structure) it is possible to infer the number of suitable clusters. Hierarchical clustering generally falls into two types: agglomerative, with a bottom-up approach, in this case, all elements start separately and are grouped in stages, one by one to form clusters, and the divisive, with a top-down approach where all elements start together in a single cluster and divisions are performed recursively as the hierarchy is descended. As with the agglomerative method, we choose the optimal number of clusters from all possible combinations.

Guan et al. (2018), in his study, applied hierarchical clustering to explore food choices at meals in a sample of overweight and obese participants. This algorithm allowed the identification of food clusters closely related to meals based on reported foods and item frequencies in the screening of dietary data. According to the authors, these results can aid in the development of strategies to improve food choices and behavior change at the individual level through a deeper understanding of these choices.

Association rules algorithm. Association rules comprise another type of unsupervised learning. However, instead of clustering instances, this technique aims to discover potentially relevant associations or regularities between items (or attribute values) of the instances (Lakshmi and Vadivu 2017). The following implication can represent rules: $X \rightarrow Y$, where X is called rule antecedent and Y is called consequent.

The most well-known association rule algorithm is Apriori. Initially, it identifies the frequent individual items, i.e., those whose the number of occurrences in the dataset is greater or equal to a threshold called minimum support. In the second iteration, the algorithm seeks for frequent pair of items containing the frequent individual items of the previous iteration, taking the same minimum support into account. Similarly, in the third iteration, the algorithm determines the frequent item triplets containing the pairs found in iteration 2. Apriori keeps augmenting the sets of frequent items in each following iteration until no changes are detected. Finally, the resulting frequent item sets are used to build the association rules that unveil trends in the dataset. To evaluate the putative rules, a minimum confidence value is considered. The confidence value measures the force of the implication described by the rule, i.e., it measures how often items in Y appear in instances that contain X. Rules generated with a confidence value below the minimum are discarded. Additionally, the lift measure can be used to evaluate the degree of correlation between X and Y in a rule (Lakshmi and Vadivu 2017).

The study by Jiang et al. (2020) used Apriori algorithm to reveal correlations between dietary factors and anthropometric changes in middle-aged Danish citizens. This study successfully identified subgroups that shared similar dietary, lifestyle, and anthropometric profiles. The authors mention that this algorithm effectively contributes to the evaluation of eating habits assessed by food frequency questionnaires, and that was able to retrieve known association rules, such as the beneficial role of fruits and red meats in relation to changes in waist circumference in both sexes.

As demonstrated in this review, supervised learning algorithms, whether of classification or regression, are more widely used for data mining. This is because supervised learning is a much more objective task when compared to unsupervised learning which has an exploratory characteristic. Additionally, for the same reason, supervised learning models are easier to validate and there are more validation metrics available. Also, the possibility of classifying future instances with the resulting model is of broad application (Tan, Steinbach, and Kumar 2006).

On the other hand, it is not always possible to perform supervised learning as it is common the situation in which only unlabeled data is available. In such cases, unsupervised learning can be of great utility and that is why this machine learning field is also of great importance (Tan, Steinbach, and Kumar 2006). Exploratory analyses make it possible to obtain initial insights and an understanding of the behavior of the data, thus facilitating the conduct of research that does not yet have a final objective outlined.

Computational tools

The big volume of data generated in daily health, including the different types of medical records, whether electronic or

paper, has led to a change in traditional data analysis forms. Historically, the most used programs in the medical field are SAS, SPSS, and STATA. However, some difficulties may arise, such as updating or adding datasets of different types and sources or in unstructured data such as text and images (Fernandes and Filho 2019).

In this work, the change in data analysis tools was noted, as the vast majority of studies used the R software. This software is an open-source, multi-platform, and free statistical environment created by Ross Ihaka and Robert Gentleman in 1997 (Matloff 2009). R has become so popular because it has a wide variety of integrated functions, packages, and libraries that perform from simple to more complex tasks, such as applying statistical tests and ML algorithms (Murrell 2005; Matloff 2009).

Therefore, there was a growing trend to use the R software instead of classic statistical software, especially for researchers who are not in the field of computing, such as health, since R is more accepted by the scientific community and is a more complete program.

The use of WEKA in the studies included in this review is also noteworthy. Its high adoption is probably due to its simplicity, to the fact that it contains many ML libraries implemented and ready to use, and, very importantly, it can be used without prior programming knowledge.

Strengths

This systematic review had several strengths, including the fact that it is the first systematic review that analyzed the application of different ML algorithms to evaluate food intake in healthy and unhealthy individuals. In addition, we include all studies regardless of the characteristics of the populations studied, the type of study, the language, and the year of publication. This decision allowed a broader search to identify and include all studies investigating food intake with the application of different ML algorithms. Another strong point was the inclusion of studies that identified food and beverage images to evaluated food consumption. This inclusion allowed a more comprehensive review of the ML application in the area of health, focusing on nutrition, highlighting the growth in the use of these algorithms in recent years and the countries involved in these researches. Besides, both the main methods used to evaluate food intake and the main ML algorithms as well as computational tools employed were presented.

Conclusion

This review summarizes the latest information on the use of different ML algorithms to evaluate food intake. It can serve as a guide for health professionals who want to work in the area of AI. It is concluded from the results found that, currently, there is a great and growing interest in the use of ML algorithms in the area of nutrition, mainly due to a significant increase in publications in recent years.

In addition, it is also noted that the supervised learning algorithms, more precisely those based on Decision Trees, were the most used. The more frequent use of DT is possibly because they are fast to apply, simple to understand, and whose results are easy to interpret and explain. In addition, there was a change in the use of computational tools for statistical analysis, with a tendency to use other software, such as R, for being more complete, instead of the classic statistical programs.

Regarding the assessment of food intake, we observed that the FFQ was the most used method since it allows a long-term evaluation, is simple, fast, and easy to administer and process. However, even understanding the importance of investigating food intake in each population and how the use of ML algorithms can be interesting, there was little diversity of countries involved in the studies analyzed. In this sense, it is encouraged that studies focusing on the application of ML algorithms in the investigation of food intake in each country are conducted, as the problems faced by different regions require different levels of research and intervention for the development of food reeducation programs and specific public policies. Furthermore, health professionals should understand that the use of ML is a collaborative activity, combining professional experience with data analysis and processing, in order to facilitate decision making in planning and delivering health care.

We would like to stress that there are other machine learning methods applied to food intake other than the ones that we cited here. However, we exploited the most-frequently applied ML procedures to provide an overview of the main ML methods used in relevant publications in recent years.

We suggest to researchers who use machine learning techniques in their studies that they mention broader search terms – such as: machine learning, deep learning, and data mining – in their texts and not just the specific names of techniques, in order to expand their visibility and make it easier the identification of their articles during the use of search engines.

Author contributions

L.O.C., A.L.G.D., D.L.F., and J.B. designed the study. L.O.C and A.L.G.D. selected and reviewed the articles and extracted the data. L.O.C., A.L.G.D., and D.L.F. analyzed and interpreted the data and drafted the manuscript. J.B., R.D-B., and F.R.C improved the manuscript and critically revised the scientific content. All authors read and approved the final manuscript.

Conflict of interest

The authors have no relevant interests to declare.

Funding

This work was supported by the Fundação de Amparo a Pesquisa do Estado de Minas Gerais (FAPEMIG), Belo Horizonte, Brazil; the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Brasilia, Brazil; and the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Brasilia, Brazil.

ORCID

Larissa Oliveira Chaves http://orcid.org/0000-0002-6962-2284
Ana Luiza Gomes Domingos http://orcid.org/0000-0001-7010-0574
Daniel Louzada Fernandes http://orcid.org/0000-0002-6548-294X
Fabio Ribeiro Cerqueira http://orcid.org/0000-0003-1325-2592
Rodrigo Siqueira-Batista http://orcid.org/0000-0002-3661-1570
Josefina Bressan http://orcid.org/0000-0002-4993-9436

References

Ahn, J. S., D. W. Kim, J. Kim, H. Park, and J. E. Lee. 2019. Development of a smartphone application for dietary self-monitoring. *Frontiers in nutrition* 6:1–12. doi:10.3389/fnut.2019.00149.

Al-Maqaleh, B. M., and A. M. G. Abdullah. 2017. Intelligent predictive system using classification techniques for heart disease diagnosis. *International Journal of Computer Science Engineering (IJCSE)* 6 (6): 145–51.

Assari, R. P., Azimi, and M. R. Taghva. 2017. Heart disease diagnosis using data mining techniques. *International Journal of Economics & Management Sciences* 06 (03):1–5. doi:10.4172/2162-6359.1000415.

Babu, S., E. M. Vivek, K. P. Famina, K. Fida, P. Aswathi, M. Shanid, and M. Hena. 2017. Heart disease diagnosis using data mining technique. *Electronics, Communication and Aerospace Technology (ICECA), International Conference* 1:750–3.

Bodnar, L. M., A. R. Cartus, S. I. Kirkpatrick, K. P. Himes, E. H. Kennedy, H. N. Simhan, W. A. Grobman, J. Y. Duffy, R. M. Silver, S. Parry, et al. 2020. Machine learning as a strategy to account for dietary synergy: An illustration based on dietary intake and adverse pregnancy outcomes. *The American Journal of Clinical Nutrition* 111 (6):1235–43. doi:10.1093/ajcn/nqaa027.

Burgermaster, M., J. H. Son, P. G. Davidson, A. M. Smaldone, G. Kuperman, D. J. Feller, K. G. Burt, M. E. Levine, D. J. Albers, C. Weng, et al. 2020. A new approach to integrating patient-generated data with expert knowledge for personalized goal setting: A pilot study. *International Journal of Medical Informatics* 139:104158. doi:10.1016/j.ijmedinf.2020.104158.

Chmurzynska, A., M. A. Mlodzik-Czyzewska, A. M. Malinowska, J. Czarnocinska, and D. J. Wiebe. 2018. Use of a smartphone application can improve assessment of high-fat food consumption in overweight individuals. *Nutrients* 10 (11):1692–12. doi:10.3390/nu10111692.

ColorBrewer. 2020. Programa Color Brewer 2.0 Color Advice for cartography. Disponpivelem. Accessed 2020. https://colorbrewer2.org/.

Cutillo, C. M., K. R. Sharma, L. Foschini, S. Kundu, M. Mackintosh, and K. D. Mand. 2020. Machine intelligence in healthcare-perspectives on trustworthiness, explainability, usability, and transparency. *NPJ Digital Medicine* 3:47. doi:10.1038/s41746-020-0254-2.

Dao, M. C., A. F. Subar, M. Warthon-Medina, J. E. Cade, T. Burrows, R. K. Golley, N. G. Forouhi, M. Pearce, and B. A. Holmes. 2019. Dietary assessment toolkits: An overview. *Public Health Nutrition* 22 (3):404–418. doi:10.1017/S1368980018002951.

De Cos Juez, F. J., F. S. Lasheras, P. J. G. Nieto, and M. A. S. Suárez. 2009. A new data mining methodology applied to the modelling of the influence of diet and lifestyle on the value of bone mineral density in postmenopausal women. *International Journal of Computer Mathematics* 86 (10-11):1878–87. doi:10.1080/00207160902783557.

De Cos Juez, F. J. M A., Suárez-Suárez, F. S. Lasheras, and A. Murcia-Mazón. 2011. Application of neural networks to the study of the influence of diet and lifestyle on the value of bone mineral density in post-menopausal women. *Mathematical and Computer Modelling* 54 (7-8):1665–1670. doi:10.1016/j.mcm.2010.11.069.

Dey, A. 2016. Machine learning algorithms: A review. *International Journal of Computer Science and Information Technologies* 7 (3): 1174–1179.

Dipnall, J. F. J A., Pasco, M. Berk, L. J. Williams, S. Dodd, F. N. Jacka, and D. Meyer. 2017. Getting RID of the blues: Formulating a Risk Index for Depression (RID) using structural equation modeling. *Australian & New Zealand Journal of Psychiatry* 51 (11):1121–13. doi:10.1177/0004867417726860.

Easton, J. F. H R., Sicilia, and C. R. Stephens. 2019. Classification of diagnostic subcategories for obesity and diabetes based on eating patterns. *Nutrition & Dietetics: The Journal of the Dietitians Association of Australia* 76 (1):104–109. doi:10.1111/1747-0080.12495.

Faruqui, S. H. A. Y., Du, R. Meka, A. Alaeddini, C. Li, S. Shirinkam, and J. Wang. 2019. Development of a deep learning model for dynamic forecasting of blood glucose level for type 2 diabetes mellitus: Secondary analysis of a randomized controlled trial. *JMIR mHealth and uHealth* 7 (11):e14452. doi:10.2196/14452.

Fernandes, F. T., and A. D. P. C. Filho. 2019. Data mining and machine learning perspectives for occupational safety and health. *Revista Brasileira de Saúde Ocupacional* 44:e13. doi:10.1590/2317-6369000019418.

Forman, E. M. S P., Goldstein, R. J. Crochiere, M. L. Butryn, A. S. Juarascio, F. Z. Zhang, and G. D. Foster. 2019. Randomized controlled trial of OnTrack, a just-in-time adaptive intervention designed to enhance weight loss. *Translational Behavioral Medicine* 6:1–13.

Forman, E. M. S P., Goldstein, F. Zhang, B. C. Evans, S. M. Manasse, M. L. Butryn, A. S. Juarascio, P. Abichandani, G. J. Martin, and G. D. Foster. 2019. OnTrack: Development and feasibility of a smartphone app designed to predict and prevent dietary lapses. *Translational Behavioral Medicine* 9 (2):236–245. doi:10.1093/tbm/iby016.

Ghorbani, R., and R. Ghousi. 2019. Predictive data mining approaches in medical diagnosis: A review of some diseases prediction. *International Journal of Data and Network Science* 3:47–70. doi:10.5267/j.ijdns.2019.1.003.

Giabbanelli, P. J., and J. Adams. 2016. Identifying small groups of foods that can predict achievement of key dietary recommendations: Data mining of the uk national diet and nutrition survey, 2008-12. *Public Health Nutrition* 19 (9):1543–1551. doi:10.1017/S1368980016000185.

Goldstein, B. A., A. M. Navar, and R. E. Carter. 2017. Moving beyond regression techniques in cardiovascular risk prediction: Applying machine learning to address analytic challenges. *European Heart* 38 (23):1805–1814.

Guan, V. X. Y C., Probst, E. P. Neale, M. J. Batterham, and L. C. Tapsell. 2018. Identifying usual food choices at meals in overweight and obese study volunteers: Implications for dietary advice. *The British Journal of nutrition* 120 (4):472–480. doi:10.1017/S0007114518001587.

Hamad, R., Z. S. Templeton, L. Schoemaker, M. Zhao, and J. Bhattacharya. 2019. Comparing demographic and health characteristics of new and existing SNAP recipients: Application of a machine learning algorithm. *The American Journal of clinical nutrition* 109 (4):1164–1172. doi:10.1093/ajcn/nqy355.

He, X. B R., Matam, S. Bellary, G. Ghosh, and A. K. Chattopadhyay. 2020. CHD risk minimization through lifestyle control: Machine learning gateway. *Scientific Reports* 10 (1):4090. doi:10.1038/s41598-020-60786-w.

Hearty, A. P., and M. J. Gibney. 2008. Analysis of meal patterns with the use of supervised data mining techniques-artificial neural networks and decision trees. *The American Journal of Clinical Nutrition* 88 (6):1632–42. doi:10.3945/ajcn.2008.26619.

Higgins, J. P. T., and S. Green. 2011. *Cochrane handbook for systematic reviews of interventions*. West Sussex, England: John Wiley & Sons.

Iwendi, C. S., Khan, J. H. Anajemba, A. K. Bashir, and F. Noor. 2020. Realizing an efficient IoMT-assisted patient diet recommendation system through machine learning model. *IEEE Access* 8: 28462–28474. doi:10.1109/ACCESS.2020.2968537.

Jain, A. K. 2010. Data clustering: 50 years beyond k-means. *Pattern Recognition Letters* 31 (8):651–666. doi:10.1016/j.patrec.2009.09.011.

Jia, W. Y., Li, R. Qu, T. Baranowski, L. E. Burke, H. Zhang, Y. Bai, J. M. Mancino, G. Xu, Z.-H. Mao, et al. 2019. Automatic food detection in egocentric images using artificial intelligence technology.

Public Health Nutrition 22 (7):1168–1179. doi:10.1017/S1368980018000538.

Jiang, L., K. Audouze, J. A. R. Herrera, L. H. Angquist, S. K. Kjaerulff, J. M. G. Izarzugaza, A. Tjønneland, J. Halkjaer, K. Overvad, T. I. A. Sørensen, et al. 2020. Conflicting associations between dietary patterns and changes of anthropometric traits across subgroups of middle-aged women and men. *Clinical Nutrition (Edinburgh, Scotland)* 39 (1):265–275. doi:10.1016/j.clnu.2019.02.003.

Kan, H. J., H. Kharrazi, H.-Y. Chang, D. Bodycombe, K. Lemke, and J. P. Weiner. 2019. Exploring the use of machine learning for risk adjustment: A comparison of standard and penalized linear regression models in predicting health care costs in older adults. *PLoS One* 14 (3):e0213258. doi:10.1371/journal.pone.0213258.

Kanerva, N. J., Kontto, M. Erkkola, J. Nevalainen, and S. Mannisto. 2018. Suitability of random forest analysis for epidemiological research: Exploring sociodemographic and lifestyle-related risk factors of overweight in a cross-sectional design. *Scandinavian Journal of Public Health* 46 (5):557–564. doi:10.1177/1403494817736944.

Khan, A. B., Baharudin, L. H. Lee, and K. Khan. 2010. A review of machine learning algorithms for text-documents classification. *Journal of advances in information technology* 1 (1):4–20.

Kodati, S. R., Vivekanandam, and G. Ravi. 2019. Comparative analysis of clustering algorithms with heart disease data sets using data mining weka tool. *Soft Computing and Signal Processing* 111–117.

Kwon, Y-JH S., Kim, D.-H. Jung, and J.-K. Kim. 2020. Cluster analysis of nutritional factors associated with low muscle mass index in middle-aged and older adults. *Clinical Nutrition (Edinburgh, Scotland)* 39 (11):3369–3376. doi:10.1016/j.clnu.2020.02.024.

Lakshmi, K. S., and G. Vadivu. 2017. Extracting association rules from medical health records using multi-criteria decision analysis. *Procedia Computer Science* 115:290–95.

Latha, R., and T. Thegaleesan. 2019. Complexity of food choice and statistical techniques. *International Journal of Innovative Studies in Sociology and Humanities* 4 (2):90–95.

Lazarou, C., M. Karaolis, A.-L. Matalas, and D. B. Panagiotakos. 2012. Dietary patterns analysis using data mining method. An application to data from the CYKIDS study. *Computer Methods and programs in biomedicine* 108 (2):706–714. doi:10.1016/j.cmpb.2011.12.011.

Liberati, A. D G., Altman, J. Tetzlaff, C. Mulrow, P. C. Gøtzsche, J. P. A. Ioannidis, M. Clarke, P. J. Devereaux, J. Kleijnen, and D. Moher. 2009. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: Explanation and elaboration. *Research Methods & Reporting* 339:b2700. doi:10.1136/bmj.b2700.

Lundberg, S. M., and S.-I. Lee. 2017. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems* 4765–4774.

Ma, S., and X. Chen. 2019. A data mining approach to predict risk of cardiovascular. *AIP Conference Proceedings*.

Matloff, N. 2009. *The art of R programming: A tour of statistical software design*. 373 p.

Mezgec, S., and B. K. Seljak. 2017. NutriNet: A deep learning food and drink image recognition system for dietary assessment. *Nutrients* 9 (7):657. doi:10.3390/nu9070657.

Michalski, R. S., J. G. Carbonell, and T. M. Mitchell. 2013. *Machine learning: An artificial intelligence approach*. Springer Science & Business Media.

Murrell, P. 2005. *R graphics*. 1st ed. Editora: Chapman and Hall/CRC; 328 p.

Mutter, S. A. E., Casey, S. Zhen, Z. Shi, and V.-P. Mäkinen. 2017. Multivariable analysis of nutritional and socio-economic profiles shows differences in incident anemia for Northern and Southern Jiangsu in China. *Nutrients* 9 (10):1153. doi:10.3390/nu9101153.

Narziev, N. H., Goh, K. Toshnazarov, S. A. Lee, K.-M. Chung, and Y. Noh. 2020. STDD: Short-term depression detection with passive sensing. *Sensors* 20 (5):1396. doi:10.3390/s20051396.

Ordóñez, C. J M., Matías, J. F. De Cos Juez, and P. J. García. 2009. Machine learning techniques applied to the determination of osteoporosis incidence in post-menopausal women. *Mathematical and Computer Modelling* 50 (5-6):673–679. doi:10.1016/j.mcm.2008.12.024.

Pagamunici, L., M. De Souza, A. H. P. Gohara, A. K. Silvestre, A. A. F. Visentainer, J. V. De Souza, N. E. Gomes, and S. T. M. Matsushita. 2014. Multivariate study and regression analysis of gluten-free granola. *Food Science and Technology* 34 (1):127–134. doi:10.1590/S0101-20612014005000005.

Panaretos, D. E., Koloverou, A. C. Dimopoulos, G.-M. Kouli, M. Vamvakari, G. Tzavelas, C. Pitsavos, and D. B. Panagiotakos. 2018. A comparison of statistical and machine-learning techniques in evaluating the association between dietary patterns and 10-year cardiometabolic risk (2002-2012): The ATTICA study. *The British Journal of Nutrition* 120 (3):326–334. doi:10.1017/S0007114518001150.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, and V. Dubourg. 2011. Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research* 12:2825–2830.

Popkin, B. M. L S., Adair, and S. W. Ng. 2012. Global nutrition transition and the pandemic of obesity in developing countries. *Nutrition Reviews* 70 (1):3–21. doi:10.1111/j.1753-4887.2011.00456.x.

Rajput, A. R P., Aharwal, M. Dubey, S. Saxena, and M. Raghuvanshi. 2011. J48 and JRIP rules for e-governance data. *International Journal of Computer Science and Security* 5 (2):201–207.

Reis, R. H., Peixoto, J. Machado, and A. Abelha. 2017. Machine learning in nutritional follow-up research. *Open Computer Science* 7 (1): 41–45. doi:10.1515/comp-2017-0008.

Rodriguez-Galiano, V., M. Sanchez-Castillo, M. Chica-Olmo, and M. Chica-Rivas. 2015. Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines. *Ore Geology Reviews* 71:804–818. doi:10.1016/j.oregeorev.2015.01.001.

Rosso, N., and P. Giabbanelli. 2018. Accurately inferring compliance to five major food guidelines through simplified surveys: Applying data mining to the UK National Diet and Nutrition Survey. *JMIR Public Health and surveillance* 4 (2):e56. doi:10.2196/publichealth.9536.

Rupasinghe, W. S. W. A., H. T. S. Perera, and N. M. J. Wickramaratne. 2020. A comprehensive review on dietary assessment methods in epidemiological research. *Public Health Nutrition* 3 (1):204–211.

Savage, A. H., Bambrick, and D. Gallegos. 2020. From garden to store: Local perspectives of changing food and nutrition security in a Pacific Island country. *Food Security* 12 (6):1331–1348.

Shao, Z., C. Chen, W. Li, H. Ren, and W. Chen. 2019. Assessment of the risk factors in the daily life of stroke patients based on an optimized decision tree. *Technology and health care: Official journal of the European Society for Engineering and Medicine* 27 (S1):317–S329. doi:10.3233/THC-199030.

Sharp, D. B., and M. Allman-Farinelli. 2014. The feasibility and validity of mobile phones to assess dietary intake. *Nutrition (Burbank, Los Angeles County, Calif.)* 30 (11-12):1257–1266. doi:10.1016/j.nut.2014.02.020.

Shiao, S. P. K., J. Grayson, A. Lie, and C. H. Yu. 2018a. Personalized nutrition—Genes, diet, and related interactive parameters as predictors of cancer in multiethnic colorectal cancer families. *Nutrients* 10 (6):795. doi:10.3390/nu10060795.

Shiao, S. P. K., J. Grayson, A. Lie, and C. H. Yu. 2018b. Predictors of the healthy eating index and glycemic index in multi-ethnic colorectal cancer families. *Nutrients* 10 (6):674. doi:10.3390/nu10060674.

Shim, J.-S., K. Oh, and H. C. Kim. 2014. Dietary assessment methods in epidemiologic studies. *Epidemiology and health* 36:e 2014009. doi:10.4178/epih/e2014009.

Shiokawa, Y., Y. Date, and J. Kikuchi. 2018. Application of kernel principal component analysis and computational machine learning to exploration of metabolites strongly associated with diet. *Scientific Reports* 8 (1):3426. doi:10.1038/s41598-018-20121-w.

Silva, B. V. R., M. G. Rad, J. Cui, M. Mccabe, and K. Pan. 2018. A mobile-based diet monitoring system for obesity management. *Journal of Health & Medical Informatics* 9 (2):1–20.

Silvera, S. A. N., S. T. Mayne, M. D. Gammon, T. L. Vaughan, W.-H. Chow, J. A. Dubin, R. Dubrow, J. L. Stanford, A. B. West, H. Rotterdam, et al. 2014. Diet and lifestyle factors and risk of subtypes

of esophageal and gastric cancers: Classification tree analysis. *Annals of epidemiology* 24 (1):50–57. doi:10.1016/j.annepidem.2013.10.009.

Singh, P. S., Singh, and G. S. Pandi-Jai. 2018. Effective heart disease prediction system using data mining techniques. *International Journal of nanomedicine* 13:121–124. doi:10.2147/IJN.S124998.

Siqueira-Batista, R., and E. Silva. 2019. Notas sobre os fundamentos matemáticos da Inteligência Artificial. *Revista De Ciência, Tecnologia e Inovação* 4:44–54.

Smallwood, R. D., and E. J. Sondik. 1973. The optimal control of partially observable Markov processes over a finite horizon. *Operations Research* 21 (5):1071–1088. doi:10.1287/opre.21.5.1071.

Tan, P. N., M. Steinbach, and V. Kumar. 2006. *Introduction to data mining*. São Carlos: Pearson Education.

Vasileska, A., and G. Rechkoska. 2012. Global and regional food consumption patterns and trends. *Procedia - Social and Behavioral Sciences* 44:363–369. doi:10.1016/j.sbspro.2012.05.040.

Vucic, V., M. Glibetic, R. Novakovic, J. Ngo, D. Ristic-Medic, J. Tepsic, M. Ranic, L. Serra-Majem, and M. Gurinovic. 2009. Dietary assessment methods used for low-income populations in food consumption surveys: A literature review. *British Journal of Nutrition* 101 (S2):S95–S101. doi:10.1017/S0007114509990626.

Xu, R., B. E. Blanchard, J. M. McCaffrey, S. Woolley, L. M. L. Corso, and V. B. Duffy. 2020. Food liking-based diet quality indexes (DQI) generated by conceptual and machine learning explained variability in cardiometabolic risk factors in young adults. *Nutrients* 12 (4):882. doi:10.3390/nu12040882.

Yu, E. Y. W., A. Wesselius, C. Sinhart, A. Wolk, M. C. Stern, X. Jiang, L. Tang, J. Marshall, E. Kellen, P. van den Brandt, et al. 2020. A data mining approach to investigate food groups related to incidence of bladder cancer in the bladder cancer epidemiology and nutritional determinants international study. *The British Journal of nutrition* 124 (6):611–619. doi:10.1017/S0007114520001439.

Zeevi, D., T. Korem, N. Zmora, D. Israeli, D. Rothschild, A. Weinberger, O. Ben-Yacov, D. Lador, T. Avnit-Sagi, M. Lotan-Pompan, et al. 2015. Personalized nutrition by prediction of glycemic responses. *Cell* 163 (5):1079–1094. doi:10.1016/j.cell.2015.11.001.

Zenitani, S. H., Nishiuchi, and T. Kiuchi. 2010. Smart-card-based automatic meal record system intervention tool for analysis using data mining approach. *Nutrition Research (New York, N.Y.)* 30 (4): 261–270. doi:10.1016/j.nutres.2010.04.003.

Zheng, Q., H. Delingette, K. Fung, S. E. Petersen, and N. Ayache. 2019. Unsupervised shape and motion analysis of 3822 cardiac 4D MRI of UK Biobank. *Preprint submitted toarXiv.*

5.2 **Artigo 2:** Cardiovascular disease analysis using unsupervised machine learning approach: Brazilian Cardioprotective Nutritional Program (BALANCE program)

Cardiovascular disease analysis using unsupervised machine learning approach: Brazilian Cardioprotective Nutritional Program (BALANCE program)

Larissa Oliveira Chaves, MD[a,*], Daniel Louzada Fernandes, MD[b], Ana Luiza Gomes Domingos PhD[a], Fabio Ribeiro Cerqueira, PhD[b,c], Rodrigo Siqueira-Batista, PhD[d], Ângela Cristine Bersh-Ferreira, PhD[e], Camila Ragne Torreglosa PhD[e], Aline Marcadenti, PhD[e,f], Bernardete Weber, PhD[e], Josefina Bressan, PhD[a]

[a]Department of Nutrition and Health, Universidade Federal de Viçosa, Viçosa – Minas Gerais, Brazil;

[b]Department of Computer Science, Universidade Federal de Viçosa, Viçosa – Minas Gerais, Brazil;

[c]Department of Production Engineering, Universidade Federal Fluminense, Petrópolis – Rio de Janeiro, Brazil;

[d]Department of Medicine and Nursing, Universidade Federal de Viçosa, Viçosa – Minas Gerais and School of Medicine of the Faculdade Dinâmica do Vale do Piranga, Ponte Nova – Minas Gerais, Brazil ;

[e]Research Institute, Hospital do Coração, São Paulo – São Paulo, Brazil;

[f]Graduate Program in Health Sciences (Cardiology), Institute of Cardiology/Fundação Universitária de Cardiologia do Rio Grande do Sul, Porto Alegre – Rio Grande do Sul, Brazil.

Funding sources: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

*Corresponding author: Larissa Oliveira Chaves, Department of Nutrition and Health, Universidade Federal de Viçosa, Avenida PH Rolfs s/n, Viçosa, Minas Gerais, CEP 36570-900, Brazil. Telephone: +5531-9-92862908. Email: larissa.chaves@ufv.br, larissaochaves@yahoo.com.br

Abstract

Cardiovascular diseases (CVD) are the main causes of death worldwide and have a financial impact on the quality of life of individuals. Consequently, machine learning (ML) techniques in the health field have been used for disease prevention and prediction. This work aims to analyze data from a population with CVD, using unsupervised ML methods, to identify groups of multivariate profiles and describe their characteristics. This is a cross-sectional study with baseline data from the "Brazilian Cardioprotective Nutritional Program - BALANCE Program" that included 1.990 patients with socioeconomic, clinical, biochemical, and behavioral characteristics. A grouping procedure was performed using the K-means algorithm. Four distinct clusters were found, two composed exclusively by men and two by women. Higher male income was identified in the male clusters; a higher education level; a greater number of smokers and ex-smokers; more physical activity practitioners; higher consumption of calories, polyunsaturated fatty acids, and omega 6; lower consumption of saturated fatty acids. In addition, a more frequent presence of coronary artery disease and acute myocardial infarction was observed. The clusters composed of women showed more obesity, diabetes mellitus, systemic arterial hypertension, dyslipidemia, and more risk factors for CVD. The results revealed differences in sex and the use of hypoglycemic agents in CVD, which can lead to several cardiovascular events. We show that ML techniques can be a robust tool to explore patterns and relationships applied to the CVD problem to guide different treatment approaches and public policies.

Keywords: Chronic non-communicable disease, cardiovascular disease, machine learning, clustering techniques, *k*-means algorithm.

Introduction

Cardiovascular Diseases (CVD) belong to the group of Chronic Non-communicable Diseases (NCD) responsible for the main causes of death globally. They include coronary, cerebrovascular, rheumatic heart disease and other conditions.[1]It is estimated that 17.9 million people die each year from these diseases, representing 31% of all deaths globally, with 75% occurring in low and middle-income countries and one third in people under 70 years of age. [1] The increasing occurrence of those diseases is known to be related to various risk factors, such as age, sex, family history of CVD, genetic predisposition, smoking, physical inactivity, inadequate diet, insomnia, diabetes mellitus, systemic arterial hypertension (SAH), hypercholesterolemia, and obesity.[2] CVD is not only a public health

problem but also an economic challenge. A European study by the European Heart Network estimated that in 2015 CVD costed the economy 210 billion euros, out of which 53% were due to health costs, 26% due to productivity losses, and 21% due to health care for people with CVD.[3] In Brazil, the cost of hospitalizations for CVD patients is considered the highest among the causes of hospital admissions, 88% with medication, 66% with social security, and 33% with morbidity.[4] Considering the need to improve the quality of health services and reduce costs, computational procedures have been proposed, mainly based on Machine Learning (ML) techniques, to understand large sets of data that are generated daily in the health field to assist professionals in the decision-making process.[5] ML is a subarea of Artificial Intelligence whose objective is to enable the discovery of knowledge in an automatic or semi-automatic way through algorithms that detect and extract patterns in datasets quickly and accurately, with a broad application on several fields, such as space technology, statistics, bioinformatics, economics, business, among others.[6,7] ML has been widely used in the area of health. Studies highlight the importance of applying algorithms to extract knowledge from data, aiming to prevent and predict diseases, early diagnosis, effective treatments, and reduction of costs and hospital deaths.[8-11] Specifically, unsupervised ML techniques for clustering analysis - that is the focus of this work - have been employed in health care to assess groups of patients with SAH,[12] with two or more NCD,[13] and organizing physical activity training for sports teams.[14] This demonstrates that clustering algorithms can accurately identify homogeneous groups with specific characteristics to assist in pattern detection. Studies emphasizing health data analysis using ML techniques are necessary to understand hidden correlations in these data and propose effective treatments and public policies. In this sense, the objective of this study was to perform an exploratory analysis of data from a population with established CVD through clustering algorithms and data visualization methods to identify distinct multivariate profile groups and investigate their similarities and differences to guide appropriate treatment approaches and public policies.

Methods

This is a cross-sectional study with baseline data from the multicenter study: “Brazilian Cardioprotective Nutritional Program (BALANCE Program)” registered on ClinicalTrials.gov (NCT01620398), coordinated by the *Hospital do Coração* (HCor) and in partnership with the Brazilian Ministry of Health.

Since this is a multicenter study, each Center submitted its study protocol to the local Ethics Committee, and the studies were initiated after all protocols were approved.[15]All patients included in the study signed a consent form. The study protocol was developed following the Declaration of Helsinki and the Brazilian and international ethical principles.[16]

Participants of both sexes, 45 years of age or older, with current evidence or in the last ten years of at least one CVD, were included. The following criteria were adopted for the confirmation of CVD: (i) Coronary Arterial Disease (CAD) presence of one or more symptoms: asymptomatic, symptomatic or treated CAD and Acute Myocardial Infarction (AMI); (ii) Cerebrovascular Disease: Cerebral Vascular Accident (CVA), Transitory Ischemic Attack (TIA) and Encephalic Vascular Accident (AVE); (iii) Peripheral Arterial Disease (PAD) presence of one or more symptoms: Asymptomatic, symptomatic or treated PAD, amputation, and aortic aneurysm. A medical report has confirmed the presence of the diseases. All eligibility criteria are reported in the study protocol.[15]

The following data were obtained through questionnaires applied by trained interviewers: (i) socioeconomic conditions: sex, age, family income and level of education; (ii) behavior: physical activity, smoking habit, and exposure to smokers; (iii) history of diseases: SAH, diabetes, dyslipidemia and family history of CAD; and (iv) use of drugs[15].

The weight, height, and waist circumference (WC) - assessed from the midpoint between the lower edge of the costal arch and the iliac crest in the middle axillary line [17] - were measured by pairs of interviewers, using the mean of each pair of measurements according to standardized methods. The Body Mass Index (BMI) was calculated by weight (kg)/height (m)² to assess the nutritional status of adults and the elderly.[18,19] The waist-to-height ratio (WHtR) was calculated by WC (cm) and height (cm) as an indicator of central obesity and was considered altered when the result was ≥ 0,5.[20]

Systolic Blood Pressure (SBP) and Diastolic Blood Pressure (DBP) were measured by trained health professionals, using the mean of two measurements, with a mercury sphygmomanometer, following the recommendations of the American Heart Association.[21] Blood samples were collected after fasting for 12 to 14 hours. Classic cardiovascular risk markers such as Triglycerides (TG), Total Cholesterol (TC), fasting glycemia, and High-Density Lipoprotein (HDL) were measured by the enzymatic colorimetric method (Johnsons & Johnsons, Raritan, EUA, VITROS 5600) and Low-Density Lipoprotein (LDL), determined by the Friedewald equation. Insulin resistance was estimated by the Triglyceride-Glucose Index (TyG Index), calculated by the formula below (1).[22] Visceral

adiposity index (VAI) was calculated by the formulas below for men (2) and women (3) to estimate visceral adiposity dysfunctions associated with cardiometabolic risk.[23]

$$Ln\left[\frac{\text{fasting TG } (mg/dl) * \text{fasting glucose } (mg/dl)}{2}\right] \quad (1)$$

$$\left[\frac{WC\ (cm)}{(39{,}69+1{,}88 * BMI(kg/m2))}\right] * \left(\frac{TG\ (mmol/L)}{1{,}03}\right) * \left(\frac{1{,}31}{HDL\ (mmol/L)}\right) \quad (2)$$

$$\left[\frac{WC\ (cm)}{(36{,}58+1{,}89 * BMI\ (kg/m2))}\right] * \left(\frac{TG\ (mmol/L)}{0{,}81}\right) * \left(\frac{1{,}52}{HDL\ (mmol/L)}\right) \quad (3)$$

The cardiometabolic risk factors considered were: obesity, physical inactivity, smoking, WC, artery pressure, TC, LDL, HDL, TG, glycemia, classified according to the Brazilian Dyslipidemias and Atherosclerosis Prevention Guideline update.[24]

Dietary intake was assessed by the average food consumption recorded by two 24-hour recalls (R24H) and the nutrient intake estimated by the Nutriquanti® computer program.[25] Each nutrient was adjusted to 1.000 kcal of energy intake. Based on the information provided by R24H, food and preparations were also classified through NOVA. NOVA is a Brazilian classification that gathers foods into four groups according to the processing (in natura or minimally processed foods; processed culinary ingredients; processed foods; and ultra-processed foods) to which these foods are submitted.[26]The mixed preparations were classified according to the proportion of the main ingredients. The dietary intake profile was expressed according to the caloric contribution of grouped foods in relation to the daily energy intake adjusted to 1.000 kcal of energy intake.

To apply ML techniques to the data collection, the development environment Jupyter Notebook (version 6.0.3)[27] and the programming language Python (version 3.7.6).[28]The following data science libraries were used in our Python scripts: pandas and numpy for data manipulation, matplotlib for data visualization, scipy for statistics, and sklearn for the ML algorithms.

To improve the visual analysis of the results, the webColorBrewer 2.0 tool[29]was used to choose the best color palette to highlight the clusters found and such that colorblind individuals can differentiate those clusters.

Data preprocessing

This phase is considered one of the most important of the whole knowledge discovery process since redundant attributes or any other inconsistency in the data may influence the detection of relevant information. Therefore, proper conduct of this phase can improve the performance of ML algorithms.[30]

Initially, the obtained dataset included 2.535 patients (instances) and 207 variables (attributes). A first analysis conducted by the team of researchers (nutritionists, physicians, and computer scientists) let to the manual removal of 81 attributes and 545 instances from the dataset. The criteria used for removal were clearly redundancy (e.g., the same measure is encompassed by two variables, one as a percentage value and another as an absolute value) and missing values that could not be imputed, as it compromises the quality of the data.

A descriptive analysis of the data based on mean, standard deviation, extreme values, percentiles, and quartiles was subsequently performed. Then, the dispersion, variance, and outliers were analyzed graphically using histograms and boxplots for proper visualization and initial understanding of basic and fundamental features of the data. Based on the obtained data summary, 64 more attributes could be removed. As a result, the final dataset ended up with 62 attributes and 1.990 instances.

To avoid scale problems in the calculation of distance between instances – which is a fundamental step in clustering procedures – the attributes were normalized. Numerical attribute values were normalized to the range [0, 1], while nominal attribute values were mapped to ordinal values as bins between 0 and 1. Thus, for example, binary attribute values were mapped to 0 and 1; ternary attribute values were mapped to 0, 0.5, 1, and so on.

The dataset was once more analyzed aiming to decrease the number of attributes, but this time using an algorithm for dimensionality reduction, namely, Principal Component Analysis (PCA).[31] It is known that the decrease in dimensionality reduces the complexity of the data, which often makes algorithms run faster and produce better results.[32] PCA is an orthogonal linear transformation of the data to a new coordinate system in a way that the greatest variance lies on the first coordinate (called the first principal component), the second greatest variance lies on the second coordinate (called the second principal component). After a series of experiments to define which components to preserve, only the first and the second principal components were kept, comprising 47.27% of variation in the data. The criterion used to choose the components was to maximize the silhouette coefficient (see sections below) in the clustering procedure. Therefore, even though the variation

captured by the chosen components was only 47.27%, the resulting clusters obtained in the following stages presented higher quality than in the cases that more components were considered.

Data processing

After preparing the dataset, ML algorithms can then be applied to perform the required data mining task. However, there is no previous knowledge about correlations, patterns, or tendencies in the obtained dataset in this study. Most importantly, there is no previously defined class of interest. Therefore, we chose an unsupervised ML approach, specifically the application of clustering techniques, to perform an exploratory analysis of the resultant data.

Clustering algorithms are used to explore unlabeled data, i.e., datasets whose instances are not previously associated with any category (or class). Note that when the class values are known, classification algorithms (supervised ML) can be applied instead to find attribute relations to predict the class of future unknown instances. In the case of clustering, on the other hand, as no class is associated with the instances, the goal is to find cohesive and well-separated groups – that can be thought of as hidden categories – and then study each group (more commonly referred to a cluster) separately to reveal previously unknown information.

A fundamental step to establish the clusters is the calculation of proximity between the instances, i.e., to measure how similar or dissimilar the instances are to each other. There are many alternatives to perform such a calculation, depending mostly on the type of attributes. However, awell-known and general method is the Euclidean distance,[33] the proximity measurement chosen in our work.

The following clustering algorithms were evaluated in our experiments: *K*-means,[34,35] hierarchical agglomerative clustering,[13,36] expectation-maximization,[33,37] and spectral clustering.[38] A key decision, in this case, is the determination of the number of clusters that these algorithms have to produce. A simple and effective approach is to test several values and calculate quality measurement – such as the silhouette coefficient – for each resulting clustering to compare all obtained clusterings. Based on this general idea, the elbow method was applied in the case of *K*-means, hierarchical clustering, and spectral clustering.[39-41] For the expectation-maximization algorithm, we used the convergence of the statistical methods Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC).[33,39,42] In all cases, the experiments pointed out that a suitable number of clusters in the given dataset would be four.

After defining the number of clusters, the next step was to compare the quality of the four clusters delivered by each tested algorithm. For this purpose, we used the silhouette coefficient, which combines measurements of both cohesion and separation, i.e., its value denotes how closely related are instances within a cluster and how well-separated the clusters are from each other.[28,34,35] The silhouette coefficient varies in the range [-1, 1]. Values close to 1 indicate a high-quality clustering, i.e., satisfactory cohesion and separation, values near zero mean indifferent clusters, i.e., the distance between clusters is not significant, and negative values indicate that a number of instances may have been assigned to the wrong clusters.[28,34,45]

Interestingly, the four algorithms produced the same clusters, thus presenting the same clustering silhouette coefficient value: 0.898, indicating a high-quality clustering. As the algorithms reported the same output, we chose *K*-means to proceed with the experiments due to its simplicity, faster execution time, and wide application in biomedical studies.

The *K*-means algorithm is presented below.[34] Note that the value for *K*, i.e., the number of clusters, must be specified in advance. As described above, in this work $K = 4$. It can be seen in the algorithm that each cluster is associated with a centroid (center point). The centroid is typically the mean of the points in the cluster. However, the initial centroids (step 1) are often chosen randomly. As already mentioned, the proximity in our case is measured by the Euclidean distance.

Algorithm *K*-means

1: Select *K* points as the initial centroids.

2: **repeat**

3: Form *K* clusters by assigning all points to the closest centroid.

4: Recompute the centroid of each cluster.

5: **until** The centroids do not change

To further highlight the quality of the clustering returned by the *K*-means algorithm, the resulting silhouette and scatter plots are shown in Figure 1, in which the found clusters are colored accordingly.[34,35] As can be seen, the silhouette plot shows a significant number of instances with silhouette values higher than the mean silhouette value that represents the clustering, which is illustrated by the dashed red line. Also, in the silhouette plot, the clusters

present similar sizes, and no cluster presented a band (tail) below zero, which would mean instances associated with a wrong cluster.[34,35] The scatter plot built using the first and the second principal components, in turn, depicts clearly well-separated clusters.

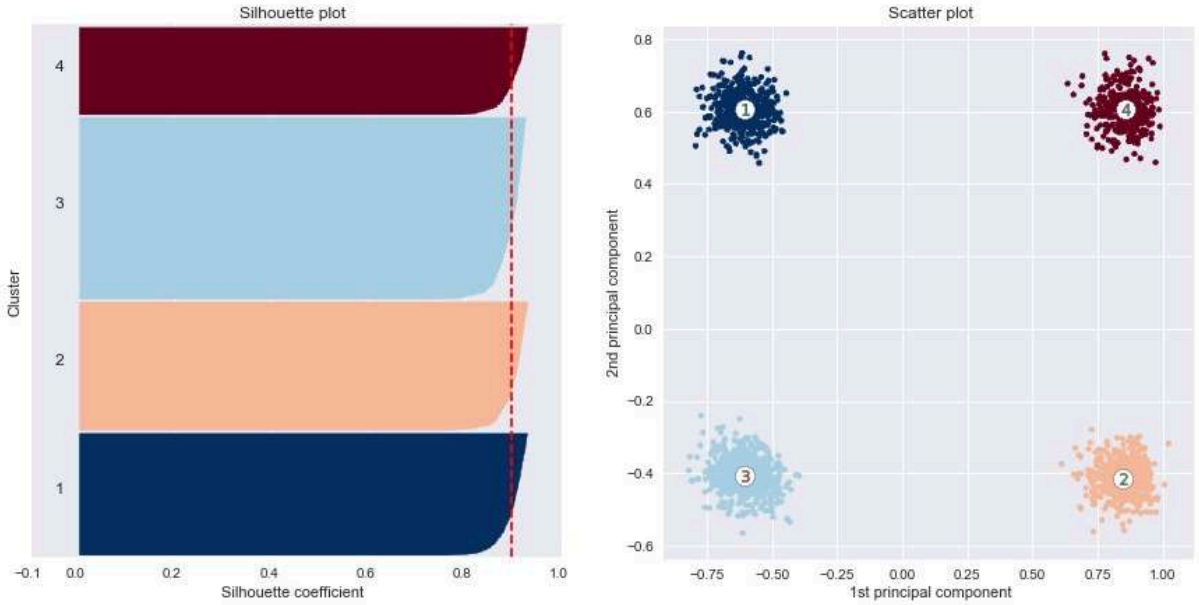


Figure 1: Quality assessment of the clustering found by the *K*-means algorithm. The left side shows the silhouette coefficient plot, while the right side shows a scatter plot of the first and the second principal components.

Data postprocessing

After defining a robust clustering, the last stage in the data analysis procedure is the study and comparison of the resulting clusters aiming for pattern and knowledge extraction. As a first approach in this vein, the results were represented, analyzed, and interpreted using parallel coordinate plots,[43] as shown in Figures 2, 3, 4, and 5. These plots depict, respectively, socioeconomic data, clinical and biochemical data, behavioral characteristics, cardiovascular events, and risk factors for CVD. A parallel coordinate plot allows multivariate data analysis in high dimensionality by drawing *n* lines for *n* clusters. The lines represent the mean values of each attribute in each cluster and will enable the comparison of attribute values across distinct groups. The y axis values are normalized on a scale of 0 to 1.

Standard statistical tests were also applied in a second approach for comparing the content of the obtained clusters. First, the Shapiro-Wilk test was applied to test the normality of the attributes, and it was observed that they did not present a normal distribution. Therefore, the statistical comparisons of the attributes across the clusters were performed using non-parametric tests. For the numeric attributes, we used the Kruskal-Wallis test,

followed by the Conover post-hoc test. Note that, in this case, the median (interquartile interval) is used in the tests instead of the mean. Categorical attributes, in turn, were evaluated by the Pearson's Chi-square test and the Fisher's Exact test. The statistical analyses were also performed using Python. For all comparative analyses, a p-value < 0.05 was considered statistically significant.

Results

After the use of the described ML techniques, the patients were categorized into cluster 1 (waxy blue),cluster 2 (salmon), cluster 3 (sky blue), and cluster 4 (venetian red). Of the 1.990 participants included in this analysis, 58.4% were male and 63 years old on average (57-70). The majority of individuals had family income between 4 and 10 minimum salaries. In addition, it was observed that 53.8% were ex-smokers, 65% did not practice physical activity, 89.8% had hypertension, 78.7% had dyslipidemia, 43% had diabetes mellitus, and 65.9% had a family history of CAD. As for the use of medications, 90% used anticoagulants or antiplatelet and antihypertensives (Table 1).

Table 1: Characterization of the population with cardiovascular diseases according to the categorical variables.

Variables	**Total (*n* = 1990)**	**Group 1 (*n* = 466)**	**Group 2 (*n*= 493)**	**Group 3 (*n* =697)**	**Group 4 (*n* = 344)**	***p*values**
Sex						**< 0.001**
Female	827 (41.56)	0	493 (59.61)	0	334 (40.39)	
Male	1.163 (58.44)	466 (40.07)	0	697 (59.93)	0	
Level of education						**< 0.001**
Illiterate/incomplete fundamental	558 (28.04)	105 (18.82)	154 (27.60)	172 (30.82)	127 (22.76)	
Fundamental I	652 (32.76)	142 (21.78)	175 (26.84)	216 (33.13)	119 (18.25)	
Fundamental II	262 (13.17)	65 (24.81)	61 (23.28)	94 (35.88)	42 (16.03)	
High school	360 (18.09)	101 (28.06)	80 (22.22)	144 (40.00)	35 (9.72)	
Superior	158 (7.94)	53 (33.54)	23 (14.56)	71 (44.94)	11 (6.96)	
Family income						**< 0.001**
≥10 minimum salaries	269 (13.52)	31 (11.52)	94 (34.94)	81 (30.11)	63 (23.42)	
≥4 e <10 minimum salaries	1.139 (57,24)	257 (22.56)	305 (26.78)	368 (32.31)	209 (18.35)	
>1 e ≤3 minimum salaries	535 (26.88)	163 (30.47)	88 (16.45)	223 (41.68)	61 (11.40)	
≤1 minimum salaries	47 (2.36)	15 (31.91)	6 (12.77)	25 (53.19)	1 (2.13)	
Smoking habit						**< 0.001**
Never smoked	773 (38.84)	142 (18.37	246 (31.82)	201 (26.00)	184 (23.80)	
Smoker	145 (7.29)	30 (20.69)	45 (31.03)	50 (34.48)	20 (13.79)	
Ex-smokers	1.072 (53.87)	294 (27.43)	202 (18.84)	446 (41.6)	130 (12.13)	

Exposure to smoke						**< 0.001**
Never	1.328 (66.73)	324 (24.40)	326 (24.55)	449 (33.81)	229 (17.24)	
1 time/week	65 (3.27)	12 (18.46)	19 (29.23)	27 (41.54)	7 (10.77)	
2 times/week	160 (8.04)	30 (18.75)	34 (21.25)	73 (45.63)	23 (14.37)	
3-6 times/week	48 (2.41)	17 (35.42)	6 (12.50)	21 (43.75)	4 (8.33)	
Daily	389 (19.55)	83 (21.34)	108 (24.76)	127 (32.65)	71 (18.25)	
Physical Activity						**< 0.001**
Practitioner	695 (34.92)	174 (25.04)	147 (21.15)	291 (41.87)	83 (11.94)	
Non-practitioner	1.295 (65.08)	292 (22.55)	346 (26.72)	406 (31.35)	251 (19.38)	
Presence of diseases						
SAH	1.787 (89.80)	433 (24.23)	441 (24.68)	586 (32.79)	327 (18.30)	**< 0.001**
Diabetes	857 (43.07)	439 (51.23)	47 (5.48)	47 (5.48)	324 (37.81)	**< 0.001**
Dyslipidemia	1.568 (78.79)	394 (25.13)	366 (23.34)	521 (33.23)	287 (18.30)	**< 0.001**
Obesity	753 (37.84)	175 (23.24)	194 (25.76)	204(27.09)	180 (23.90)	**< 0.001**
CAD Family History	1.313 (65.98)	280 (21.33)	355 (27.04)	424 (32.29)	254 (19.35)	**< 0.001**
Use of medication						
Anticoagulants or antiplatelet agents	1.800 (90.45)	438 (24.33)	433 (24.06)	632 (35.11)	297 (16.50)	**0.007**
Antihypertensives	1.887 (94.82)	452 (23.95)	463 (24.54)	650 (34.45)	322 (17.06)	**0.013**
Hypoglycemic	800 (40.20)	466 (58.25)	0	0	334 (41.75)	**< 0.001**
Cardiovascular events						
Asymptomatic CAD	328 (16.48)	79 (24.09)	84 (25.61)	109 (33.23)	56 (17.07)	0.899
Symptomatic CAD	712 (35.78)	150 (21.07)	178 (25.00)	255 (35.81)	129 (18.12)	0.258
Treated CAD	1.376 (69.15)	345 (25.07)	295 (21.44)	521 (37.86)	215 (15.63)	**< 0.001**
AMI	982 (49.35)	228 (23.22)	222 (22.61)	378 (38.49)	154 (15.68)	**0.008**
CVA	235 (11.81)	54 (22.98)	66 (28.09)	82 (34.89)	33 (14.04)	0.502
Asymptomatic PAD	83 (4.17)	18 (21.69)	28 (33.73)	22 (26.51)	15 (18.07)	0.190
Symptomatic PAD	136 (6.83)	37 (27.21)	33 (24.26)	34 (25.00)	32 (23.53)	**0.027**
Treated PAD	72 (3.62)	15 (20.83)	17 (23.61)	23 (31.94)	17 (23.61)	0.486
Amputation	31 (1.56)	5 (16.13)	9 (29.03)	11 (35.48)	6 (19.35)	0.766
Aortic aneurysm	39 (1.96)	4 (10.26)	15 (38.46)	18 (46.15)	2 (5.13)	**0.011**
Number of risk factors						**< 0.001**
1-5	681 (34.22)	98 (14.39)	187 (27.46)	356 (52.28)	40 (5.87)	
≥ 6	1.309 (65.78)	368 (28.11)	306 (23.38)	341 (26.05)	294 (22.46)	

Pearson's Chi-square and Fisher's Exact Test. Data are shown as frequencies (%).

Statistically significant values (p-value < 0.05).

Abbreviations: SAH - Systemic Arterial Hypertension; CAD - Coronary Arterial Disease; AMI - Acute Myocardial Infarction; CVA - Cerebral Vascular Accident; PAD - Peripheral Arterial Disease.

Minimum salaries (R$724.00 in 2014).

According to the median of clinical, anthropometric, and biochemical attributes, it was observed that most of the individuals were pre-obese evaluated by BMI, pre-diabetic according to blood glycemia values, present insulin resistance according to TyG index, and had increased risk of metabolic complications assessed by WC, WHtR and VAI (Table 2). Regarding cardiovascular events, 40.5% of the patients presented only one event, 34%

presented two events, and 25.5% presented three or more events, with the highest prevalence being treated CAD (69.2%) and AMI (49,4%).

Table 2: Characterization of the population with cardiovascular diseases according to anthropometric, clinical, biochemical, and dietary intake data.

Variables	**Total (*n* = 1990)**	**Group 1 (*n* = 466)**	**Group 2 (*n* = 493)**	**Group 3 (*n* = 697)**	**Group 4 (*n*= 334)**	***p*value**
Age (years)	63 (57-70)	63 (57-70)	63 (58-70)	63 (57-70)	62 (56-69)	0.453
Weight (kg)	74.90 (65.12-84.67)	79.40 (71.00-89.20)[a]	68.20 (59.70-76.60)[b]	77.50 (69.20-86.00)[c]	73.10 (62.52-84.75)[d]	**< 0.001**
BMI (kg/m^2)	28.45 (25.68-31.79)	28.61 (25.95-32.03)[a]	28.46 (25.35-31.85)[a]	27.70(25.33-30.52)[b]	30.55 (26.75-34.50)[c]	**< 0.001**
DBP (mmHg)	80.00 (70.00-86.00)	80.00 (70.00-83.00)	80.00 (70.00-87.00)	80.00 (70.00-85.00)	80.00 (70.00-90.00)	0.140
SBP (mmHg)	130.00 (120.00-140.00)	130.00 (120.00-140.00)[a,b]	130.00 (120.00-140.00)[a]	126.00 (116.00-140.00)[b]	130.00 (120.00-150.00)[c]	**< 0.001**
WC(cm)	100.00 (92.00-107.00)	103.00 (95.00-110.00)[a]	95.00 (87.00-103.00)[b]	99.00 (92.00-106.00)[c]	102.00 (94.00-110.00)[a]	**< 0.001**
Cholesterol Total (mg/dl)	162.00 (13.008-193.00)	152.00 (130.00-176.00)[a]	179.00 (152.00-209.00)[b]	158.00 (136.00-184.00)[c]	168.50 (141.00-198.50)[d]	**< 0.001**
TG (mg/dl)	137.00 (99.00-190.00)	148.50 (106.00-202.75)[a]	137.00 (98.00-177.00)[b,c]	128.00 (94.00-186.00)[c]	142.00 (105.25-201.00)[a,b]	**< 0.001**
Glycemia (mg/dl)	103.00 (92.00-124.00)	124.00 (105.00-156.00)[a]	97.00 (89.00-107.00)[b]	97.00 (89.00-105.00)[b]	122.50 (101.00-157.00)[a]	**< 0.001**
HDL (mg/dl)	41.00 (35.00-49.00)	37.00 (32.00-44.00)[a]	47.00 (39.00-55.00)[b]	39.00 (33.00-47.00)[c]	43.50 (37.00-50.00)[d]	**< 0.001**
LDL (mg/dl)	88.00 (70.00-113.00)	78.00 (61.00-100.00)[a]	102.00 (80.00-129.00)[b]	87.00 (71.00-107.00)[c]	89.50 (70.00-113.75)[c]	**< 0.001**
WHtR	0.61 (0.56-0.66)	0.61 (0.57-0.65)[a]	0.61 (0.56-0.66)[a]	0.59 (0.55-0.63)[b]	0.65 (0.60-0.71)[c]	**< 0.001**
TyG index	8.89 (8.52-9.33)	9.16 (8.71-9.57)[a]	8.78 (8.44-9.11)[b]	8.73 (8.40-9.14)[b]	9.09 (8.69-9.59)[a]	**< 0.001**
VAI	2.36 (1.54-3.52)	2.42 (1.58-3.52)[a]	2.49 (1.66-3.64)[a]	2.00 (1.28-3.03)[b]	2.88 (1.91-4.57)[c]	**< 0.001**
Minimally processed (kcal)	930.77 (688.34-1216.85)	1014.60 (752.54-1231.80)[a]	820.78 (623.32-1074)[b]	1058.92 (795.48-1396.57)[c]	758.61 (588.33-10.24)[b]	**< 0.001**
Processed (kcal)	99.00 (0.00-221.00)	139.00 (0.00-262.25)[a]	74.00 (0.00-171.00)[b]	139.00 (0.00-256.00)[a]	74.00 (0.00-148.50)[b]	**< 0.001**
Ultra-processed (kcal)	213.34 (95.83-388.21)	238.20 (111.12-401.41)	215.97 (96.01-360.32)	214.21 (94.75-412.20)	186.50 (90.12-351.00)	0.142
Culinary ingredients (kcal)	0.00 (0.00-40.00)	0.00 (0.00-30.00)[a]	0.00 (0.00-50.00)[b]	4.00 (0.00-64.00)[b]	0.00 (0.00-14.50)[a]	**<0.001**
Calories	1423.60	1446.53	1245.80 (985.09-	1513.16	1119.50	**< 0.001**

	(1071.02-1819.76)	(1201.22-1810.87)[a]	1524.93)[b]	(1215.44-1897.76)[c]	(891.35-1418.72)[d]	
Carbohydrate (g)	188.40 (167.31-210.00)	185.57 (162.33-204.61)[a]	191.80 (171.72-210.24)[b]	189.99 (166.09-215.71)[a,b]	186.26 (169.85-202.59)[a,b]	**0.007**
Fatty acids (g)	44.24 (37.34-51.02)	45.02 (38.73-52.63)[a]	44.19 (37.80-50.10)[a]	42.87 (34.51-50.26)[b]	45.67 (39.90-51.68)[a]	**< 0.001**
Protein (g)	67.96 (56.66-81.49)	69.39 (57.98-84.77)[a]	66.15 (55.64-76.80)[b]	68.58 (55.61-82.56)[a,b]	67.45 (57.65-78.02)[a,b]	**0.008**
SFA (g)	14.44 (11.58-17.30)	14.41 (11.70-17.60)[a]	15.04 (12.36-17.66)[a]	13.61 (10.49-16.75)[b]	14.95 (12.56-17.44)[a]	**< 0.001**
MUFA (g)	12.96 (10.39-15.49)	13.60 (10.84-16.39)[a]	12.92 (10.55-15.17)[a,b]	12.38 (9.61-15.37)[b]	13.09 (11.03-15.06)[a,b]	**0.001**
PUFA (g)	10.20 (8.34-12.57)	10.96 (8.73-13.41)[a]	10.06 (8.45-11.99)[b]	10.01 (7.84-12.85)[b]	10.07 (8.44-12.04)[b]	**0.003**
TRANS (g)	0.03 (0.02-0.05)	0.03 (0.02-0.05)[a]	0.03 (0.02-0.04)[b]	0.03 (0.02-0.05)[a]	0.03 (0.02-0.04)[b]	**< 0.001**
Cholesterol (mg)	179.74 (133.48-246.54)	179.16 (130.70-253.24)	182.90 (142.40-241.77)	178.81 (126.46-246.57)	177.17 (141.90-226.38)	0.810
Sodium (mg)	2684.74 (2286.64-3124.61)	2815.28 (2382.10-3297.52)[a]	2542.87 (2203.31-2928.62)[b]	2747.59 (2309.40-3205.43)[a,c]	2642.30 (2320.55-3065.00)[c]	**< 0.001**
Dietary Fiber (g)	18.41 (13.78-23.92)	19.83 (14.21-25.26)[a]	17.30 (13.03-22.30)[b]	17.89 (12.78-24.72)[b,c]	18.41 (15.01-23.37)[a,c]	**< 0.001**
Calcium (mg)	459.03 (316.35-649.01)	440.11 (313.06-641.68)[a]	496.96 (336.40-654.38)[b]	439.01 (279.68-626.72)[a]	485.50 (371.44-658.87[b]	**< 0.001**
Iron (mg)	6.26 (5.19-7.74)	6.63 (5.51-8.11)[a]	5.96 (4.93-7.11)[b]	6.31 (5.12-7.98)[c]	6.14 (5.25-7.56)[b,c]	**< 0.001**
Potassium (mg)	2008.97 (1684.01-2386.02)	2053.38 (1756.92-24.03.68)[a]	1968.48 (1651.80-2273.25)[b]	1978.66-1628.84-2401.06)[a,b]	2069.15 (1729.99-2415.95)[a,b]	**0.005**
Magnesium (mg)	179.45 (151.71-216.60)	183.63 (154.99-223.47)[a]	178.28 (151.43-209.84)[a,b]	174.59 (146.91-213.66)[b]	185.58 (158.25-218.58)[a]	**0.003**
Phosphorus (mg)	838.56 (718.59-993.63)	842.74 (724.18-1019.23)	833.83 (734.71-969.63)	838.39 (698.75-991.72)	849.67 (718.63-997.44)	0.319
Copper (mg)	0.87 (0.58-1.24)	0.82 (0.51-1.15)[a]	0.91 (0.64-1.33)[b]	0.79 (0.48-1.22)[a]	0.96 (0.75-1.29)[b]	**<0.001**
Zinc (mg)	7.67 (5.99-10.36)	8.31 (6.26-11.39)[a]	7.34 (5.84-9.81)[b]	7.50 (5.82-10.53)[b]	7.59 (6.42-10.01)[a,b]	**0.001**
Selenium (µg)	18.57 (11.55-28.34)	19.02 (11.30-30.14)	18.45 (12.01-26.32)	19.48 (11.20-30.36)	17.81 (11.97-25.10)	0.291
Omega 3 (g)	0.95 (0.67-1.19)	0.98 (0.69-1.24)[a]	0.93 (0.71-1.17)[a,b]	0.89 (0.59-1.17)[b]	0.99 (0.79-1.19)[a]	**< 0.001**
Omega 6 (g)	8.33 (6.55-10.36)	8.84 (6.77-11.01)[a]	8.11 (6.56-9.86)[b]	8.36 (6.42-10.70)[b]	8.09 (6.57-9.77)[b]	**< 0.001**
Vitamin C (mg)	76.19 (32.43-167.69)	73.57 (29.31-164.00)[a,b]	84.82 (37.07-175.30)[b,c]	61.34 (25.05-165.87)[a]	95.29 (44.16-171.41)[c]	**< 0.001**
Vitamin E (RE)	1.50 (1.10-2.18)	1.57 (1.14-2.36)	1.46 (1.09-2.10)	1.45 (1.06-2.24)	1.52 (1.17-2.12)	0.146

Vitamin B2 (mg)	0.87 (0.61-1.21)	0.87 (0.58-1.22)	0.91 (0.66-1.21)	0.85 (0.59-1.23)	0.84 (0.61-1.13)	0.125
Vitamin B12 (mg)	0.62 (0.20-1.16)	0.54 (0.05-1.08)[a]	0.77 (0.37-1.27)[b]	0.52 (0.11-1.10)[a]	0.76 (0.45-1.14)[b]	**< 0.001**
Vitamin A (U)	310.13 (171.60-513.72)	291.21 (147.87-514.97)[a,b]	321.22 (211.81-501.64)[b,c]	276.12 (125.43-490.51)[a]	352.40 (241.68-564.44)[c]	**< 0.001**
Vitamin B3 (mg)	12.40 (8.42-18.91)	12.37 (8.10-18.71)	12.08 (8.60-18.11)	12.59 (8.29-19.76)	12.53 (9.09-18.51)	0.596
Vitamin B6 (mg)	0.48 (0.33-0.71)	0.51 (0.33-0.78)[a]	0.45 (0.33-0.65)[b]	0.48 (0.31-0.74)[a,b]	0.48 (0.35-0.65)[a,b]	**0.034**
Vitamin B1 (mg)	0.70 (0.54-0.97)	0.72 (0.54-0.94)	0.69 (0.54-0.98)	0.71 (0.54-0.97)	0.70 (0.57-1.02)	0.672
Vitamin D (µg)	4.40 (0.35-30.33	3.07 (0.89-16.95)[a]	5.57 (1.81-39.75)[b]	3.47 (0.86-31.13)[a]	5.20 (2.48-26.54)[b]	**< 0.001**
Vitamin B7 (mg)	0.01 (0.00-0.04)	0.00 (0.00-0.03)[a]	0.02 (0.01-0.05)[b]	0.01 (0.00-0.03)[a]	0.04 (0.01-0.06)[c]	**< 0.001**
Vitamina B5 (mg)	0.52 (0.24-1.02)	0.43 (0.20-0.95)[a]	0.59 (0.30-1.09)[b]	0.51 (0.20-0.99)[a,c]	0.54 (0.31-0.98)[b,c]	**< 0.001**

Kruskal-Wallis test and Conover post-hoc test for comparison between groups. Presented in median (interquartile range). Statistically significant values (p-value< 0.05). Legend: Each nutrient was adjusted to 1.000 kcal of energy intake. Abbreviations: BMI - Body Mass Index; DBP - Diastolic Blood Pressure; SBP - Systolic Blood Pressure; WC - Waist Circumference; TG - Triglycerides; HDL - High Density Lipoprotein; LDL - Low Density Lipoprotein; WHtR - Waist-to-Height Ratio; TyG index - Triglyceride-Glucose Index; VAI - Visceral Adiposity Index; SFA - Saturated Fatty Acid; MUFA - Monounsaturated Fatty Acids; PUFA - Polyunsaturated Fatty Acid; TRANS - TRANS Fatty Acid.

According to socioeconomic characteristics, it was observed that clusters 1 and 3 are made up only of men with higher income and levels of education. In contrast, clusters 2 and 4 are made up only of women with lower income and levels of education (Table 1 and Figure 2).

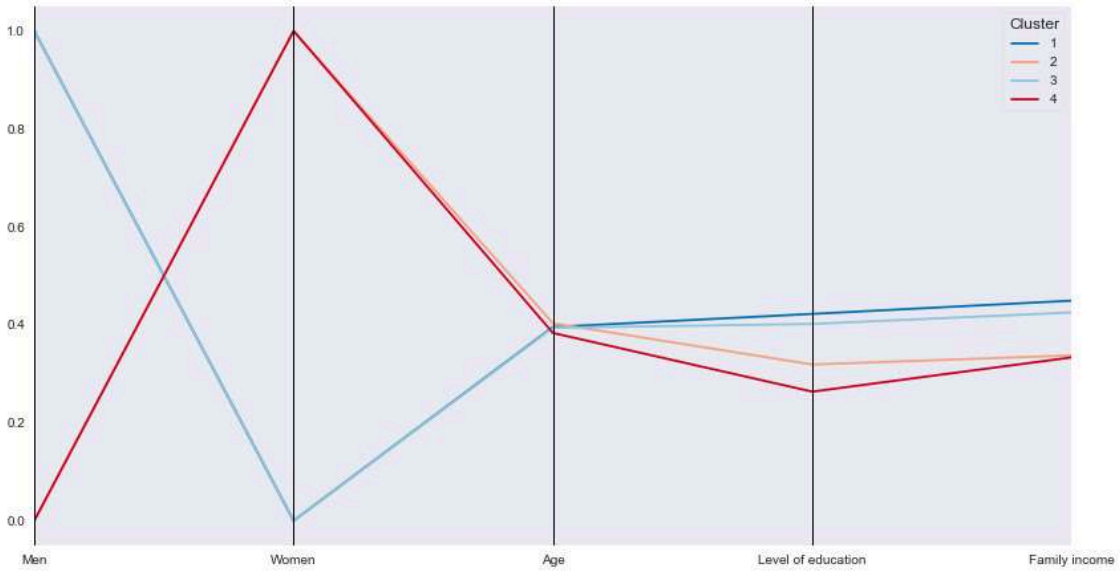


Figure 2: Parallel coordinates of socioeconomic characteristics.

The patients in cluster 4 presented higher BMI, SBP, WHtR, and VAI values regarding the clinical and biochemical characteristics. Moreover, higher values of blood glucose and TyG index were observed in clusters 1 and 4. Concerning the lipid profile, higher TC, LDL, and HDL values are found in cluster 2. It was also observed that patients with lower BMI, SBP, TG, glycemia, WHtR, TyG index, and VAI values belong to cluster 3 (Table 2 and Figure 3).

The majority of individuals in the four clusters have SAH, dyslipidemia, and CAD history; however, most patients with diabetes are concentrated in clusters 1 and 4 (Table 1). Moreover, it is noted that these diseases are more concentrated in the participants of cluster 4 and to a lesser extent in cluster 3 (Figure 3). Since it is a population with established CVD and presents another NCD, anticoagulant or antiplatelet drugs and antihypertensives in all clusters are justified. In addition, it was observed that the use of hypoglycemic agents occurred only in clusters 1 and 4.

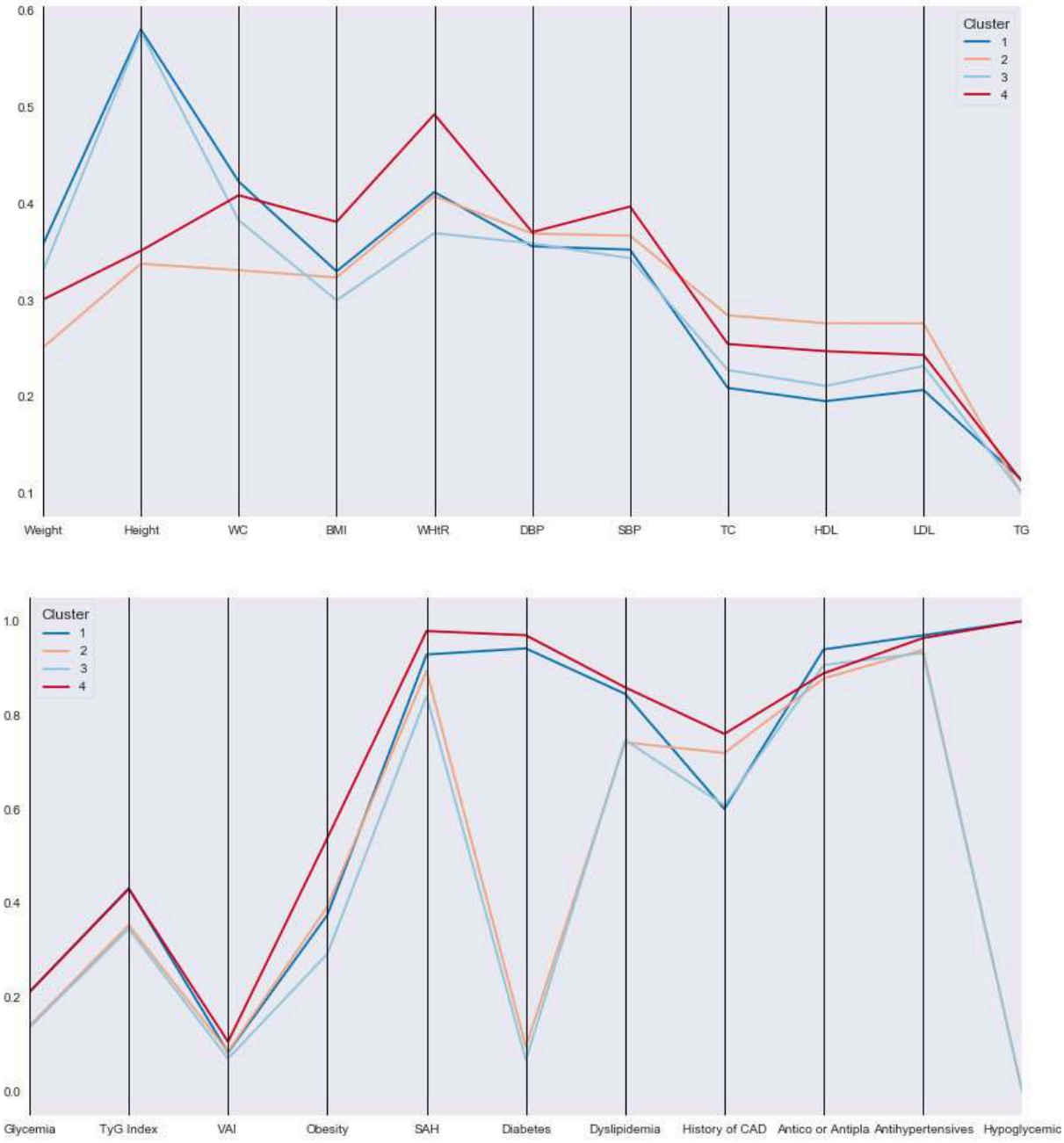
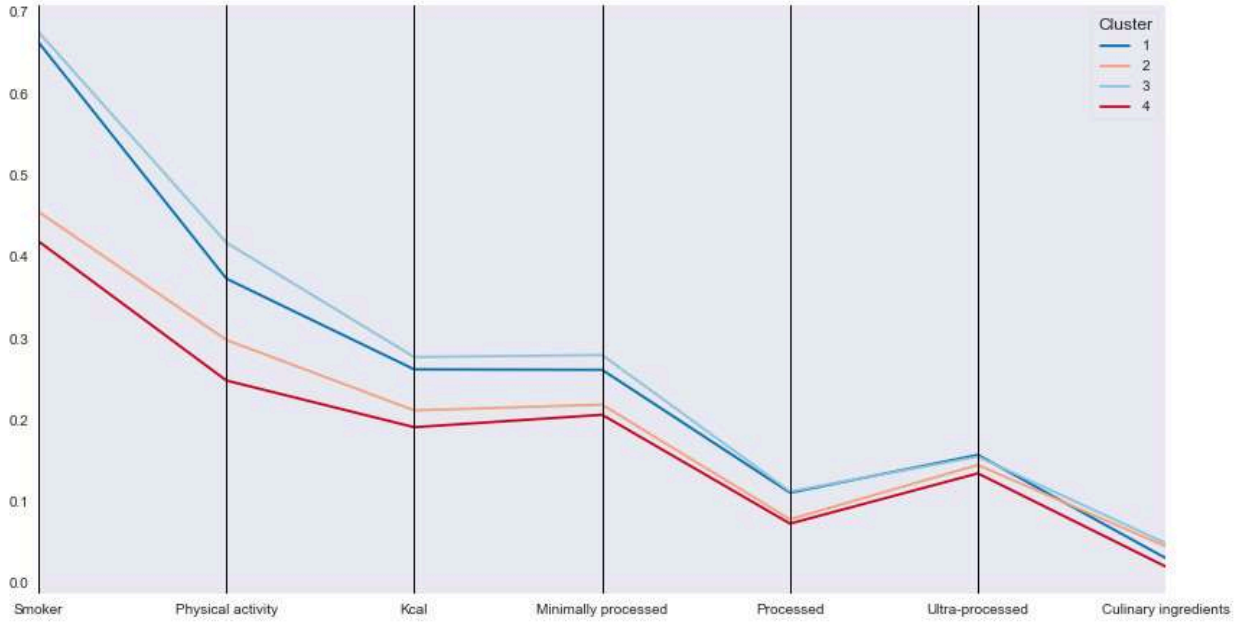


Figure 3: Parallel coordinates of clinical and biochemical characteristics. Abbreviations: WC - Waist Circumference; BMI - Body Mass Index; WHtR - Waist-to-Height Ratio; DBP - Diastolic Blood Pressure; SBP - Systolic Blood Pressure; TC - Total Cholesterol; HDL - High-Density Lipoprotein; LDL - Low-Density Lipoprotein; TG – Triglycerides; VAI - Visceral Adiposity Index; SAH - Systemic Arterial.

Regarding behavioral characteristics, smokers, ex-smokers, and those who practice physical activity are more present in clusters 1 and 3 (male) (Table 1 and Figure 4). Regarding dietary intake, higher caloric intake was observed in cluster 3 and lower intake in cluster 4. The consumption of minimally processed and processed foods was higher again in

clusters 1 and 3. Individuals in cluster 2 consumed more carbohydrates than in cluster 1, while the opposite was true for protein intake. Fatty acid total and saturated fatty acid (SFA) intake was lower in cluster 3, while polyunsaturated fatty acid (PUFA) and omega 6 intake was higher in cluster 1. Among the micronutrients, the highest intake of calcium, magnesium, and copper was in cluster 4 compared with cluster 3, and the highest intake of sodium and zinc was in cluster 1 compared with cluster 2. The highest intake of vitamins occurred in patients in clusters 3 and 4 (Vitamins A, D, and B12) and vitamins B7 and C in cluster 4 (Table 2 and Figure 4).



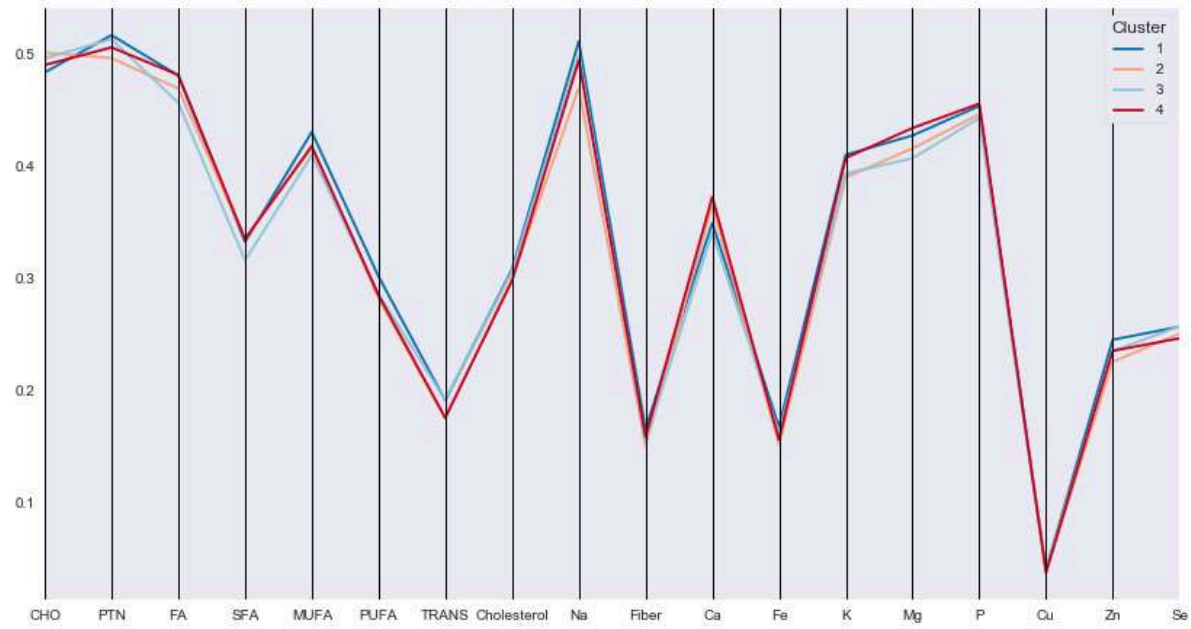


Figure 4: Parallel coordinates of behavioral characteristics.

Abbreviations: Kcal – Calories; CHO - Carbohydrate; PTN - Protein; FA – Fatty acid; SFA - Saturated Fatty Acid; MUFA - Monounsaturated Fatty Acids; PUFA - Polyunsaturated Fatty Acid; TRANS - TRANS Fatty Acid; Na - Sodium; Ca - Calcium; Fe - Iron; K - Potassium; Mg - Magnesium; P - Phosphorus; Cu - Copper; Zn - Zinc; Se – Selenium;ω−3 – Omega 3; ω−6 – Omega 6.

Regarding cardiovascular events, there were fewer cases of aortic aneurysm and amputation and more cases of CAD treated and AMI. However, more concentrated in clusters 1 and 3 (male). Most patients with symptomatic PAD were present in cluster 4 (Table 1 and Figure 5). It is also noteworthy that most patients with one to five risk factors for CVD were present in cluster 3, while the majority with more than six risk factors were in cluster 4 (Table 1 and Figure 5).

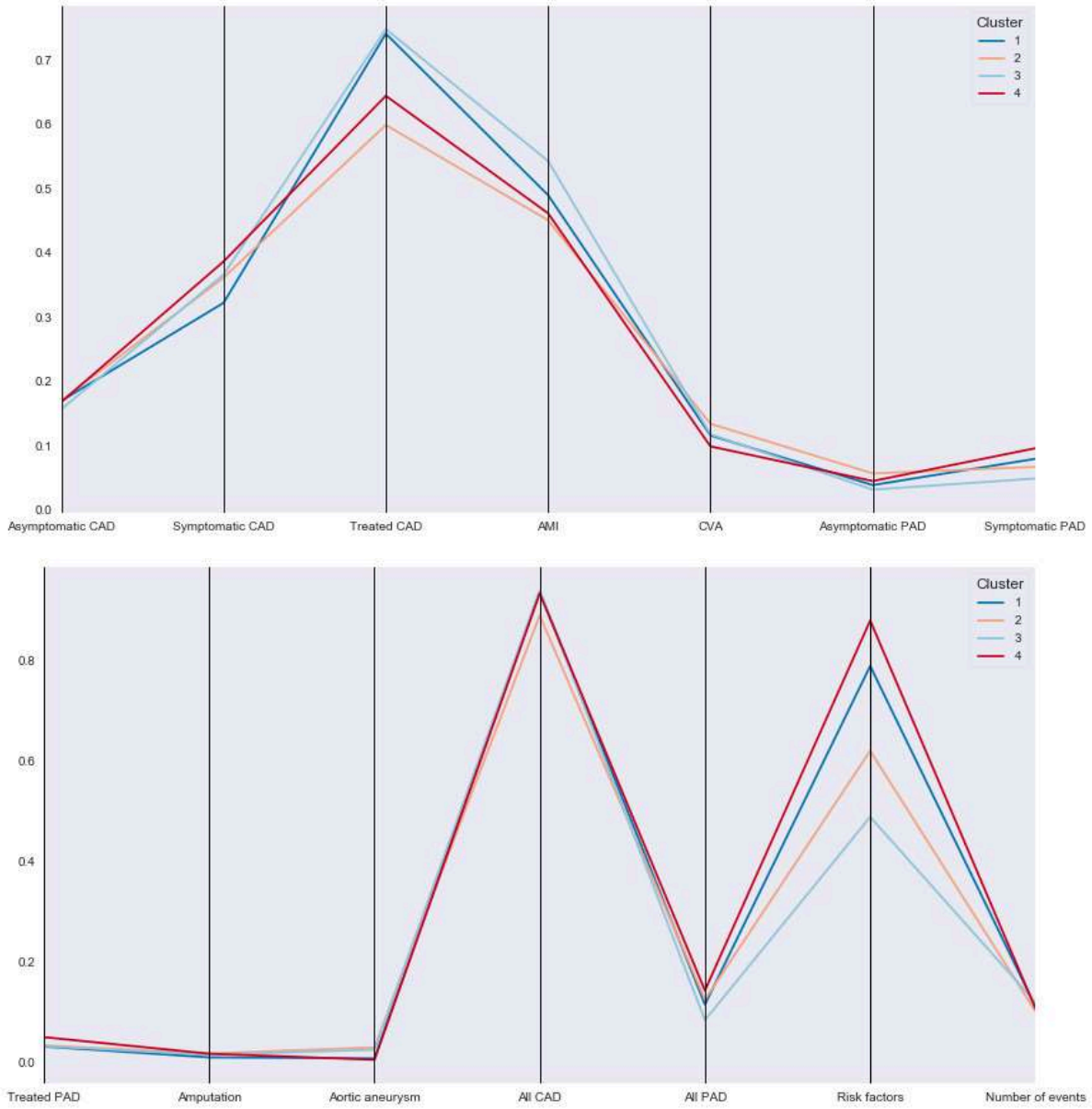


Figure 5: Parallel coordinates of cardiovascular events and risk factors for CVD.
Abbreviations: CAD - Coronary Arterial Disease; AMI - Acute Myocardial Infarction; CVA - Cerebral Vascular Accident; PAD - Peripheral Arterial Disease.

According to the results above, four clusters have different profiles in a population with established CVD. Clusters 2 and 4 – containing female participants – presented a worse profile than men, except for the consumption of nutrients. However, clusters 3 (men) and 4 (women) stood out as being the most contrasting. As can be seen in the results: cluster 3 is mostly made up of men with higher income and level of education, who practiced more physical activity, who were smokers and ex-smokers, who presented a better clinical and biochemical condition, and fewer individuals with other NCD. Cluster 4, on the other hand, is mostly made up of women with a low level of education and income, who practiced fewer

physical activities, with central and visceral obesity, diabetes mellitus, and insulin resistance, and most had other NCD and more risk factors for CVD.

Discussion

Death from CVD affects mainly individuals aged 30-69, and the forecast up to 2030 is alarming, as an increase of about 22 million deaths worldwide is projected.[44] In addition to the high impact on health system costs, CVD morbidity affects the lives of individuals and can compromise their ability to work and earn a living.[4] Therefore, the impact is not only on the health of individuals but also on health systems and the economy, which adds a huge financial burden to society, making it a social problem. The care of patients with CVD is costly, and in some countries, the financial burden caused by these diseases is expected to exceed $20 billion by 2030.[44]

It is then necessary to understand the behavior of a population with CVD to detect similarities and/or differences among individuals, which makes it possible to develop different approaches for treatments and public policies. With the help of ML, this has become possible, not to mention the importance of ML algorithms in several studies for the prediction and diagnosis of diseases.[45,46] In their review, Marques et al. describe the relevance and increasing number of initiatives that have used ML in CVD research.[47] Other studies also report that ML techniques have gained more attention and significant momentum for investigations in the cardiovascular area.[12,48]

One of the most important variables for defining the groups was sex and the use of hypoglycemic agents. This, in turn, is directly related to the diagnosis of diabetes mellitus, as seen in the results, where clusters 1 and 4 are composed of diabetic individuals and therefore used hypoglycemic agents. Regarding sex, previous studies have already pointed to differences between men and women in genetic, biological, clinical, physiological, and sociocultural terms, which strongly affect the occurrence of CVD.[49-51]These differences are highly relevant as they have an impact on pathophysiology, clinical manifestations, epidemiology, and medical approaches, and consequently the effectiveness of therapies and their possible side effects. In addition, behavioral factors such as diet and cultural aspects have a significant influence.[51] Thus, the literature shows the importance of investigating the difference between the sexes for a wider scientific understanding of the pathophysiological aspects of CVD as well as its precise diagnosis, effective treatment, and effective prevention.

This study observed important differences in socioeconomic, clinical, biochemical, and behavioral characteristics between women and men. This disparity was striking when it was found that women had lower incomes and levels of education. Some barriers make it difficult for women to access the same opportunities as men, such as the stress often associated with double occupations (work and family), economic factors, salary inequalities, and low level of education, leading to a misinformation scenario. These evidences are accompanied by higher risk and worse prognosis in women compared with men.[51,52] Thereupon, a study by Backholer et al.[53] supports the development and implementation of clinical guidelines and sex-specific policies in CVD, indicating that the importance of socioeconomic inequalities in women's cardiovascular health should be emphasized.

Smoking is known to be an important risk factor for the development of CVD.[54] The frequent reasons for this factor are increased lipid metabolism, vascular resistance, cardiac frequency, inflammation of tissues, and endothelial damage. Smoking can also reduce the oxygen supply to the heart, leading to angina, AMI, and even death.[50] Sex differences associated with smoking and CVD are increasingly frequent.[54] In Brazil, in 2019, the frequency of adult smokers was 9.8%, higher in males (12.3%) than females (7.7%). Considering the whole population, the frequency of smokers tended to be lower before the age of 25 and after the age of 65.[2] Moreover, smoking increases the risk of developing CAD from 4 to 6 times in men and from 6 to 9 times in women.[50] However, this different impact between the sexes is not fully understood, and the results are still inconsistent.

Regarding physical inactivity, it is found that the lack of exercise significantly increases the incidence of CVD.[50] A prospective study followed 73.743 postmenopausal women aged between 50 and 79 and identified that both walking and vigorous exercise could reduce the incidence of cardiovascular events.[55] Our study found that most women practiced fewer physical activities than men. Corroborating our findings, the 2011 National Health Interview Survey (NHIS) observed in adults of similar age range that physical inactivity was also higher among women (33.2%) than men (29.9%).[56] A recent study shows that in Brazil, in the year 2019, 44.8% of the population did not practice enough physical activity; this percentage being higher among women (52.2%) than men (36.1%). Moreover, the frequency of physically inactive adults was 13.9%, with minimal differences between the sexes, and in women, this frequency decreases until 54 years and increases again from that age on.[2]

Among the modifiable risk factors, adequate food consumption is considered one of the more important to prevent CVD.[57] A good diet includes the intake of fruits, vegetables,

fish, and wholefood, together with the restriction of sodium, SFA, and simple carbohydrates.[1] In this study, men in cluster 1 had a higher intake of PUFA and omega 6. In comparison, men in cluster 3 had a higher caloric intake of minimally processed, processed, and ultra-processed foods but low consumption of SFA. In 2015, in Europe, 56% of deaths from CVD in men and 48% in women were attributable to dietary factors.[3] A study conducted in Japan found that the quality of the diet is affected not only by age and sex but also by the income, occupation, and education level of individuals. The authors also suggested different approaches for improving the quality of the diet according to the family income.[58]

In Brazil, the NOVA classification has been used to assess dietary patterns according to socioeconomic and demographic distribution and associate the consumption of ultra-processed products with obesity, metabolic syndrome, and dyslipidemias.[59] In Brazilian urban households, it was noted that ultra-processed products represented 18.7% of total calories in food packages in 1988 and 29.6% in 2009, while a reduction was observed in processed food from 44% to 38.9% and in culinary ingredients from 35.7% to 28.9%.[60] It should be noted that the consumption of ultra-processed and processed is related to the increase in calorie density and SFA and simple carbohydrate intake and is associated with the decrease in dietary fiber, magnesium, and potassium, vitamin A, iron, and zinc.[60]

In our study, women also presented higher values of central and visceral obesity markers, blood pressure, and insulin resistance, and consequently more diseases associated with CVD, such as obesity, dyslipidemia, SAH, and diabetes. Studies showed that differences in sex and the presence of those diseases are higher among women.[51,54,61] Obesity, for example, is considered an independent risk factor for CVD. In the United States, two out of three adults are obese or pre-obese, and the prevalence is higher among women.[62] Concerning diabetes, from 70% to 80% of people with this disease die of CVD, and women with diabetes have a higher risk of developing cardiovascular complications.[50] In Canada and the United States, the prevalence of diabetes is higher among men than women: 7.5% vs. 5.8% in Canada and 13.6% vs. 11.2% in the US. However, the risk of developing CVD is much higher among women.[63,64] In the UK, data indicated that diabetic women were 15% less likely to complete the recommended medical care.[64]

Dyslipidemia is one of the most important risk factors for CVD, and as the population ages, the number of patients with high plasma lipid concentrations increases even more.[65] According to the study by Bello and Mosca (2004),[66] women aged between 20 and 50 have a better lipid profile than men. After menopause, however, blood lipid concentrations

increase in women, while in men of similar age the concentration measurements show stability. Studies also show that the prevalence of SAH in women before menopause is lower than that in men, but it becomes higher in women during and after menopause.[54,67] This occurs because estrogen maintains vasodilation and contributes to blood pressure control in premenopausal women.[49] Interestingly, only 23% of women and 38% of men over 80 years old can keep blood pressure < 140/90mmHg.[68] A Brazilian study observed that the prevalence of individuals over 18 years old who reported having at least one type of NCD was 44.8% – women: 50.4% and men: 39.2%. Checking the reported NCD, SAH was the most frequent in both sexes, but again women had a higher prevalence than men: 24.2% and 18.3%, respectively.[69]

As discussed above, menopause is an exclusive risk factor for CVD in women due to physiological changes in hormone levels, especially the reduction of estrogen, since it has a cardioprotective effect.[54] Estrogen has antioxidant effects, activates estrogen receptors in endothelial cells, promotes vasodilation through nitric oxide and endothelial cell growth, inhibits muscle cell proliferation, reduces LDL oxidation and platelet aggregation, and contributes to glucose homeostasis via increased glucose transport to the cell. In addition to the reduction of estrogen in menopause, excessive secretion of androgens also leads to ovarian dysfunction and insulin resistance.[70,71] Estrogen replacement is an alternative for women, but some studies have not noted a beneficial effect on reducing the incidence of CAD and, therefore, need to be better evaluated.[68]

Regarding the CVD established in this population, it was observed that the majority of male participants had CAD and AMI, while the majority of female participants developed PAD, in addition to a higher number of risk factors. Studies indicate that CAD tends to be under diagnosed in women, leading to a lower incidence, and, in contrast, tends to be more prevalent in men.[73,74] According to data from the INTERHEART study, AMI is more prevalent in women because it presents more potent risk factors such as hypertension, diabetes, and smoking.[75] About PAD is currently associated with the equivalent morbidity and mortality of CAD and AVE,[76] and a high prevalence have been observed in women under 40 and over 80 years old.[77] According to McDermott et al. (2001).[78] One of the interesting characteristics of PAD is intermittent claudication, which can occur asymptomatically or with atypical symptoms in women.

Interestingly, when investigating whether there were distinct groups in our population, we found four clusters, two consisting only of men and two only of women. This reveals

that individuals with the same chronic diseases can have important differences, which, if detected, can improve secondary prevention approaches and treatments.

The present study presents some limitations, mainly the transversal nature that does not allow the establishment of cause-effect relationships and the assessment of food consumption carried out through the R24 hours. In addition, some participants had only one R24 hours, making it impossible to carry out the average. However, the researchers involved in the data collection were trained continuously to obtain consistency in the data entered and minimize errors due to under- or overestimating food consumption. In addition, 20% of the R24 hours processed were randomly selected to evaluate data consistency. The strength of this work is that it is a multicenter study, which assessed a large number of participants with CVD in all regions of Brazil. Furthermore, data quality was maximized by automating data entry, monthly contacting the researchers, and monitoring visits to the centers.

In conclusion, our results revealed sex-related differences in CVD, whether in socioeconomic, clinical, biochemical, or behavioral aspects that could lead to different cardiovascular events. Therefore, we suggest health professionals focus on educational strategies for groups of low-income and low-education groups of women (clusters 2 and 4) and men with diabetes mellitus (cluster 1). Thus, studies on sex differences in CVD, including clinical trials to elucidate mechanisms that promote sex-specific diagnosis, treatment, and prognosis, are necessary.

Clustering analysis using unsupervised ML techniques identified four distinct profile groups, which presented significant differences, in a highly accurate manner. Therefore, it was possible to examine and describe the similarities and differences of the clusters found, which allowed us to identify important specific characteristics. Hence, it is shown again that ML techniques comprise an essential tool in the biomedical field, especially to rapidly and effectively identify patterns, facilitate knowledge extraction in massive data, and point fruitful alternatives for future research initiatives.

Although this study focused on clustering data containing CVD subjects, it is worth noting that the computational techniques employed here are generic enough to be used in other domains of interest. In particular, the use of ML techniques is highly encouraged to facilitate the acquisition of new insights that may be of fundamental importance to researchers in their investigations and health professionals in critical decision-making when it comes to health data.

Author Contributions

ACBF, CRT, and BW performed the conception and design of the study. The generation and data collection were performed by LOC, ACBF, CRT, BW, and JB. The assembly, analysis, and/or interpretation of the data were performed by LOC, DLF, ALGD, and JB. All authors contributed to drafting the manuscript, final approval for publishing submission and ensuring the integrity and accuracy of the work.

Declaration of Interests

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

Acknowledgments

We thank all the patients for participating in this project, and all participating centers: Cristiane Kovacs - Instituto Dante Pazzanese de Cardiologia, São Paulo-SP. Annie S B Moreira - Hospital Universitário Pedro Ernesto, Rio de Janeiro-RJ e Instituto Nacional de Cardiologia, Rio de Janeiro-RJ. Rosileide S Torres - Hospital das Clínicas Gaspar Vianna, Belém-PA. Helyde A Marinho - Instituto Nacional de Pesquisas da Amazônia, Manaus-AM. Cristina H de Matos - Universidade Vale do Itajaí, Itajaí-SC. Renata T A Bertacco - Universidade Federal de Pelotas, Pelotas-RS. Gabriela C Souza - Hospital de Clínicas de Porto Alegre, Porto Alegre-RS. Gabriela S Shirmann - Universidade da Região da Campanha, Bagé-RS. Francisca E Z Nagano - Hospital de Clínicas da Universidade Federal do Paraná, Curitiba-PR. Maria E M Ramos - Hospital Universitário Associação Educadora São Carlos, Canoas-RS. Soraia Poloni - Instituto de Cardiologia do Rio Grande do Sul, Porto Alegre-RS. Raquel M El Kik - Hospital São Lucas da Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre-RS. Naoel H Feres - Universidade Federal do Mato Grosso, Cuiabá-MT. Eliane S Dutra - Hospital Universitário de Brasília, Brasília-DF. Ana P P F Carvalho - Hospital das Clínicas de Goiânia, Goiânia-GO. Marta M David - Hospital Universitário Maria Aparecida Pedrossian, Campo Grande-MS. Isa G Rodrigues - Pronto Socorro Cardiológico Universitário de Pernambuco, Recife-PE. Antonio C S Sousa - Hospital São Lucas, Aracaju-SE. Amanda G L Coura - Hospital Universitário Alcides Carneiro, Campina Grande-PB. Josilene M F Pinheiro - Hospital Universitário Ana Bezerra, Santa Cruz-RN. Sandra M L Vasconcelos - Universidade Federal de Alagoas, Maceió-AL. Andreza M Penafort - Universidade de Fortaleza, Fortaleza-CE. Daniele M O Carlos -

Hospital de Messejana, Fortaleza-CE. Viviane Sahade - Hospital Universitário Professor Edgard Santos, Salvador-BA. Adriana B Luna - Hospital Universitário da Universidade Federal de Sergipe, Aracaju-SE. José A F Neto - Hospital Universitário da Universidade Federal do Maranhão, São Luís-MA. Emilio H Moriguchi - Associação Veranense de Assistência em Saúde, Veranópolis-RS. Maria C O Izar - Universidade Federal de São Paulo, São Paulo-SP. Sônia L Pinto - Universidade Federal de Tocantins, Palmas-TO. Hospital São Vicente de Paulo, Luciano M Backes - Passo Fundo-RS. Simone R Souza - Instituto Estadual de Cardiologia Aloysio de Castro, Rio de Janeiro-RJ. Magali C C - COTENUT, Porto Alegre – RS.

Disclosure statement

The authors declare no conflict of interest.

References

1. World Health Organization (WHO). Cardiovascular diseases (CVDs), 2020. Disponível em <http://origin.who.int/cardiovascular_diseases/en/>.

2. Brasil. Vigitel, Brasil 2019: vigilância de fatores de risco e proteção para doenças crônicas por inquérito telefônico. Brasília: Ministério da Saúde, 2020.

3. Wilkins E, Wilson L, Wickramasinghe K, Bhatnagar P, Leal J, Luengo-Fernandez R, Burns R, Rayner M, Townsend N. *European Cardiovascular Disease Statistics 2017. European Heart Network* 2017.

4. Siqueira ASE, Siqueira-filho AG, Land MGP. Análise do Impacto Econômico das Doenças Cardiovasculares nos Últimos Cinco Anos no Brasil. *Arquivos Brasileiros de Cardiologia*2017;109:39-46.

5. Ma S, Chen X. A Data Mining Approach to Predict Risk of Cardiovascular. *AIP Conference Proceedings* 2019;258:020014-1-020014-7.

6. Kodati S, Vivekanandam R, Ravi G. Comparative Analysis of Clustering Algorithms with Heart Disease Data sets Using Data Mining Weka Tool: Methods and Protocols. In book: Immunological Tolerance 2019:111-117.

7. Siqueira-batista R, Silva, E. Notas sobre os fundamentos matemáticos da Inteligência Artificial. *Revista De Ciência, Tecnologia e Inovação*2019;4:44-54.

8. Babu S, Vivek EM, Famina KP, Fida K, Aswathi P, Shanid M, Hena M. Heart disease diagnosis using data mining technique. *In Electronics, Communication and Aerospace Technology (ICECA), International conference* 2017;1:750-753.

9. Dekamin A, Sheibatolhamdi A. A data mining approach for coronary artery disease prediction in Iran. *Journal of Advanced Medical Science sand Applied Technologies* 2017;3:29-38.

10. Pouriyeh S, Sannino G, De Pietro G, Arabnia H, Gutierrez JB, Vahid S. A comprehensive investigation and comparison of Machine Learning Techniques in the domain of heart disease. *2017* IEEE Symposium on Computers and Communications (ISCC) 2017:204-207.

11. Singh P, Singh S, Pandi-jai GS. Effective heart disease prediction system using data mining techniques. *International Journal of Nanomedicine* 2018;13:121–124.

12. Guo Q, Lu X, Gao Y, Zhang J, Yan B, Su D, Song A, Zhao X, Wang G. Cluster analysis: a new approach for identification of underlying risk factors for coronary artery disease in essential hypertensive patients. *Scientific Reports* 2017;7:43965.

13. Newcomer SR, Steiner JF, Bayllis EA. Identifying Subgroups of Complex Patients With Cluster Analysis. *American College of Cardiology*2011;17:e324-e332.

14. Novack LF, Nascimento VB, Salgueirosa FM, Carignano LF, Fornaziero A, Gomes EB, Osieck R. Distribuição de subgrupos com base nas respostas fisiológicas em jogadores profissionais de futebol pela técnica k means cluster. *Revista Brasileira de Medicina do Esporte*2013;19:130-133.

15. Weber B, Bersch-Ferreira AC, Torreglosa CR, Ross-Fernandes MB, Silva JT, Galante AP, LaraES, Costa RP, Soares RM, CavalcantiAB, et al. The Brazilian Cardioprotective Nutritional Program to reduce events and risk factors in secondary prevention for cardiovascular disease: study protocol (The BALANCE Program Trial). *American Heart Journal* 2016;171:73–81.

16. World Medical Association. World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects. *Journal of the American Medical Association* 2013;310:2191–4.

17. World Health Organization (WHO). Waist circumference and waist–hip ratio: report of a WHO expert consultation.*WHO*, 2008.

18. World Health Organization (WHO). Physical status: the use an interpretation of anthropometry. Technical Report Series. Geneva: *WHO*, 1995.

19. OPAS – Organização Pan-Americana da Saúde. XXXVI Reunión Del Comitê Asesor de Investigaciones em Salud – EncuestraMulticêntrica – Salud Beinestar y Envejecimeiento (SABE) en América Latina e el Caribe - Informe preliminar. Disponível em <URL: //WWW.opas.org/program/sabe.htm.> 2002.

20. Ashwell M, Gunn P, Gibson S. Waist-to-height ratio is a better screening tool than waist circumference and BMI for adult cardiometabolic risk factors: systematic review and meta-analysis. *Obesity Reviews* 2012;13:275-278.

21. AHA. Understanding blood pressure Reading site. 2018.

22. Simental-mendía LE, Rodríguez-morán M, Guerrero-romero F. The product of fasting glucose and triglycerides as surrogate for identifying insulin resistance in apparently healthy subjects. *Metabolic Syndrome and Related Disorders* 2008;6:299–304.

23. Amato MC, Giordano C, Galia M, Criscimanna A, Vitabile S, Midiri A, Galluzzo A. Visceral Adiposity Index: a reliable indicator of visceral fat function associated with cardiometabolic risk. *Diabetes Care*. 2010;33:920-2.

24. Faludi AA, zar MCO, Saraiva JFK, Chacra APM, Bianco HT, Afiune NA, Bertolami A, Pereira AC, Lottenberg AM, Sposito AC, et al. Atualização da Diretriz Brasileira de Dislipidemias e Prevenção da Aterosclerose – 2017. *Arquivos Brasileiros de Cardiologia*2017;109:1-76.

25. Galante AP. Desenvolvimento e validação de um método computadorizado para avaliação do consumo alimentar, preenchido por indivíduos adultos utilizando a Web. Tese (Doutorado em Nutrição Humana Aplicada) – Faculdade de Ciências Farmacêuticas, Universidade de São Paulo, 2007.

26. Monteiro CA, Cannon G, Levy R, Moubarac J-C, Jaime P, Martins AP, Canella D, Louzada M, Parra D, Ricardo C, et al. O Sistema Alimentar. *World* 2016;7:1–3.

27. Kluyver T, Ragan-Kelley B, Pérez F, Granger BE, Bussonnier M, Frederic J, Kelley K, Hamrick JB, Grout J, Corlay S, et al. Jupyter notebooks-a publishing format for reproducible computational workflows. *In Electronic Publishing* 2016:87–90.

28. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, et al. Scikitlearn: Machine learning in python. *Journal of machine learning research* 2011:2825–2830.

29. WebColorBrewer. ProgramaColorBrewer 2.0 Color Advice for cartography. Disponpivelem<https://colorbrewer2.org/>.

30. Witten H, Frank E, Hall MA, Pal CJ. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann. 4th edition. 2016.

31. Ding C, He X. K-means clustering via principal component analysis. In Proceedings of the twenty-first international conference on Machine learning. *Association for Computing Machinery* 2004:29.

32. Cerqueira FR, Ferreira TG, OliveiraAP, Augusto DA, Krempser E, Barbosa HJC, Franceschini SCC, Freitas BACF, Gomes AP, Siqueira-Batista R. Nicesim: an open-source simulator based on machine learning techniques to support medical research on prenatal and perinatal care decision making. *Artificial intelligence in medicine* 2014;62:193–201.

33. Zheng Q, Delingette H, Fung K, Petersen SE, Ayache N. Unsupervised shape and motion analysis of 3822 cardiac 4D MRIs of UK Biobank. *Computer Vision and Pattern Recognition* 2019.

34. Jain AK. Data clustering: 50 years beyond k-means. *Pattern recognition letters* 2010;31:651–666.

35. Ahmad T, Lars H. Lund, Pooja Rao, Rohit Ghosh, Prashant Warier, Benjamin Vaccaro, Ulf Dahlström, Christopher M. O'Connor, G. Michael Felker, et al. Machine learning methods improve prognostication , identify clinically distinct phenotypes, and detect heterogeneity in response to therapy in a large cohort of heart failure patients. *Journal American Heart Association* 2018;7: e008081.

36. Brum F, Mozzaquatro PM, Zanatta JM. Estudo sobre os algoritmos de clusterização Hierarchical Clusteres e Simple K-means aplicados no agrupamento de padrões similares. *Revista da Universidade Vale do Rio Verde* 2019;17:1-9.

37. Jung YG, Kang MS, Heo J. Clustering performance comparison using *K*means and expectation maximization algorithms. *Biotechnology & Biotechnological Equipment* 2014;28:S44-S48.

38. Ding S, Cong L, Hu Q, Jia H, Shi Z. A multiway p-spectral clustering algorithm. *Knowledge-Based Systems* 2018;164:371-377.

39. Masud MA, Huang JZ, Wei C, Wang J, Khan I, Zhong M. I-nice: A new approach for identifying the number of clusters and initial cluster centres. *Information Sciences* 2018;466:129-151.

40. Hess S, Duivesteijn WK. Is the magic number - inferring the number of clusters through nonparametric concentration inequalities. *Machine Leanring and Lnowledge Discovery in Databases*. 2020:257-273.

41. Paternina MRA, Zamora-mendez A, Ortiz-Bejar J, Chow JH, Ramirez JM. Identification of coherent trajectories by modal characteristics and hierarchical agglomerative clustering. *Electric Power Systems Research* 2018;158:170-183.

42. Grall-maes E, Dao DT. Assessing the Number of Clusters in a Mixture Model with Side-information**.** *Science and Technology Publications* 2016:41-47.

43. Li M, Zhen L, Yao X. How to Read Many-Objective Solution Sets in Parallel Coordinates [Educational Forum]. *IEEE Computational Intelligence Magazine* 2017;12:88-100.

44. Patel S, Ram F, Patel SK, Kumar K. Cardiovascular diseases and health care expenditure (HCE) of inpatient and outpatient: A study from India Human Development Survey. *Clinical Epidemiology and Global Health* 2019;8:671-677.

45. Singh R, Rajesh E. Prediction of Heart Disease by Clustering and Classification Techniques. *International Journal of Computer Sciences and Engineering* 2019;7:2347-2693.

46. Wu C-S, Badshah M, Bhagwat V. Heart Disease Prediction Using Data Mining Techniques. *In Proceedings of 2019 2nd International Conference on Data Science and Information Technology(DSIT'19)* 2019:7-11.

47. Filho SEM, Fernandes FA, Soares CLA, Seixas FL, Dos Santos AASMD, Gismondi RA, Mesquita ET, Mesquita CT. Inteligência Artificial em Cardiologia: Conceitos, Ferramentas e Desafios – "Quem Corre é o Cavalo, Você Precisa ser o Jóquei". *Arquivos Brasileiros de Cardiologia*2020;114:718-725.

48. Daghistani TA, Elshawi R, SakrS, Ahmed AM, Al-Thwayee A, Al-Mallah MH.Predictors of in-hospital length of stay among cardiac patients: A machine learning approach. *International Journal of Cardiology* 2019;288:140–147.

49. Humphries KH. Izadnegahdar M, Sedlak T, SawJ, Johnston N, Schenck-GustafssonK, ShahRU, Regitz-Zagrosek V, Grewal J, Vaccarino V, et al. Sex Differences in Cardiovascular Disease – Impact on Care and Outcomes. *Front Neuroendocrinol* 2017;46:46-70.

50. Zhang Y, LiuB, Zhao R, Zhang S, Yu XY, Li Y. The Influence of Sex on Cardiac Physiology and Cardiovascular Diseases. *Journal of Cardiovascular Translational Research* 2019;13:3-13.

51. Ventura-clapier R, PiquereauJ, Garnier A, Mericskay M, Lemaire C, Crozatier B. Gender issues in cardiovascular diseases. Focus on energy metabolism. BBA -*Molecular Basis of Disease*2020;1866:165722.

52. Schultz WM, Kelli HM, Lisko JC, Varghese T, Shen J, Sandesara P, Quyyumi AA, Taylor HA, Gulati M, Harold JG, et al. Socioeconomic Status and Cardiovascular Outcomes Challenges and Interventions. *Circulation* 2018;137:2166-2178.

53. Backholer K, Peters SAE, Bots SH, Peeters A, Huxley RR, Woodward M. Sex differences in the relationship between socioeconomic status and cardiovascular disease: a systematic review and meta-analysis. *Epidemiol Community Health* 2016;71:1-8.

54. Garcia M, Mulvagh SL, Bairey Merz CN, Buring JE, Manson JE. Cardiovascular Disease in Women: Clinical Perspectives. *Circulation Research* 2016;118:1273–1293.

55. Manson JE, Greenland P, LaCroix AZ, Stefanick ML, Mouton CP, Oberman A, Perri MG, Sheps DS, Pettinger MB, Siscovick DS. Walking compared with vigorous exercise for the prevention of cardiovascular events in women. *New England Journal of Medicine* 2002;347:716–725.

56. Schiller JS, Lucas JW, Ward BW, Peregoy JA. Summary health statistics for u.S. Adults: National health interview survey 2010. *Vital Health Stat 10* 2012;252:1-207.

57. Kim J, Hoang T, Bu SY, Kim J-M, Choi J-H, Park E, Lee S-M, Park E, Min J-Y, Lee IS, et al. Associations of Dietary Intake with Cardiovascular Disease, Blood Pressure, and Lipid Profile in the Korean Population: a Systematic Review and Meta-Analysis. *Journal of Lipid Atherosclerosis* 2020;9:205-229.

58. Kurotani K, Akter S, Kashino I, Goto A, Mizoue T, Noda M, Sasazuki S, Sawada N, Tsugane S. Quality of diet and mortality among Japanese men and women: Japan Public Health Center based prospective study. *British Medical Journal* 2016;352:i1209.

59. Monteiro CA, Cannon G, Moubarac J-C, Levy RB, Louzada MLC, Jaime PC. The UN Decade of Nutrition, the NOVA food classification and the trouble with ultra-processing. *Public Health Nutrition* 2018;21:5–17.

60. Martins APB, Levy RB, Claro RM, Moubarac JC, Monteiro CA. Increased contribution of ultra-processed food products in the Brazilian diet (1987–2009). *Revista de Saúde Pública* 2013;47:656–665.

61. Wada H, Miyauchi K, Daida H. Gender differences in the clinical features and outcomes of patients with coronary artery disease**.** *Expert Review of CardiovascularTherapy* 2018;17:127-133.

62. Flegal KM, Carroll MD, Kit BK, Ogden CL. Prevalence of obesity and trends in the distribution of body mass index among us adults, 1999–2010. *Journal of the American Medical Association* 2012;307:491-497.

63. Statistics Canada. Diabetes, 2014. Retrieved from http://www.statcan.gc.ca/pub/82-625-x/2015001/article/14180-eng.htm Accessed em Dec 12, 2018. 2014.

64. National Center for Chronic Disease Prevention and Health Promotion. National Diabetes Statistics Report, 2014. Retrieved from http://www.cdc.gov/diabetes/pdfs/data/2014-report-estimates-of-diabetes-and-its-burden-in-the-united-states.pdf Accessed em Dec 12, 2018. 2014.

65. Lee M, Saver JL, Towfighi A, Chow J, Ovbiagele B. Efficacy of fibrates for cardiovascular risk reduction in persons with atherogenic dyslipidemia: a meta-analysis. *Atherosclerosis* 2011;217:492–498.

66. Bello N, Mosca L. Epidemiology of coronary heart disease in women. *Progress in Cardiovascular Diseases* 2004;46:287–295.

67. Regitz-zagrosek V, Oertelt-Prigione S, Prescott E, Franconi F, Gerdts E, Foryst-Ludwig A, Maas AHEM, Kautzky-Willer A, Knappe-Wegner D, Kintscher U, et al. Gender in cardiovascular diseases: Impact on clinical manifestations, management, and outcomes. *European Heart Journal* 2016;37:24-34.

68. Lloyd-jones DM, Evans JC, Levy D. Hypertension in adults across the age spectrum: current outcomes and control in the community. *Journal of the American Medical Association* 2005;294:466–472.

69. Malta DC, Stopa SR, Szwarcwald CL, Gomes NL, Junior JBS, Dos Reis AAC. Surveillance and monitoring of major chronic diseases in Brazil – National Health Survey, 2013. *Revista Brasileira de Epidemiologia* 2015;18:3-16.

70. Widder J, Pelzer T, Poser-klein CV, Hu K, Jazbutyte V, Fritzemeier K-H, Hegele-hartung C, Neyses L, Bauersachs J. Improvement of endothelial dysfunction by selective estrogen receptor-alpha stimulation in ovariectomized SHR. *Hypertension* 2003;42: 991–996.

71. Grady D, Herrington D, Bittner V, Blumenthal R, Davidson M, Hlatky M, Hsia J, Hulley S, Herd A, Khan S, et al. Cardiovascular disease outcomes during 6.8 years of hormone therapy: Heart and Estrogen/progestin Replacement Study follow-up (HERS II). *Journal of the American Medical Association* 2002;288:49-57.

72. Anderson GL, Limacher M, Assaf AR, Bassford T, Beresford SAA, Black H, Bonds D, Brunner R, Brzyski R, Caan B. et al. Effects of conjugated equine estrogen in

postmenopausal women with hysterectomy: the Women's Health Initiative randomized controlled trial. *Journal of the American Medical Association* 2004;291:1701–1712.

73. Pendyala LK, Torguson R, Loh JP, Kitabata H, Minha S, Badr S, Dvir D, Barbash IM, Satler LF, Pichard AD, et al. Comparison of adverse outcomes after contemporary percutaneous coronary intervention in women versus men with acute coronary syndrome. *American Journal Cardiology* 2013;111:1092-1098.

74. Toyota T, Furukawa Y, Ehara N, Funakoshi S, Morimoto T, Kaji S, Nakagawa Y, Kadota K, Iwabuchi M, Shiomi H, et al. Sex-based differences in clinical practice and outcomes for Japanese patients with acute myocardial infarction undergoing primary percutaneous coronary intervention. *Circulation Journal* 2013;77:1508–1517.

75. Walli-attaei M, Joseph P, Rosengren A, Chow CK, Rangarajan S, Lear SA, AlHabib KF, Davletov K, Dans A, Lanas F, et al. Variations Between Women and Men in Risk Factors, Treatments, Cardiovascular Disease Incidence, and Death in 27 High-Income, Middle-Income, and Low-Income Countries (PURE): A Prospective Cohort Study. *Lancet* 2020;396:97-109.

76. Mcdermott MM, Ferrucci L, Guralnik JM, Dyer AR, Liu K, Pearce WH, Clark E, Liao Y, Criqui MH. The ankle-brachial index is associated with the magnitude of impaired walking endurance among men and women with peripheral arterial disease. *Vascular Medicine* 2010;15:251–257.

77. Hirsch AT, Allison MA, Gomes AS, Corriere MA, Duval S, Ershow AG, Hiatt WR, Karas RH, Lovell MB, McDermott MM, et al. Treat-Jacobson D. A call to action: Women and peripheral artery disease: A scientific statement from the American Heart Association. *Circulation* 2012;125:1449–1472.

78. Mcdermott MM, Greenland P, Liu K, Guralnik JM, Criqui MH, Dolan NC, Chan C, Celic L, Pearce WH, Schneider JR, et al. Leg symptoms in peripheral arterial disease: Associated clinical characteristics and functional impairment. *Journal of the American Medical Association*2001;286:1599–1606.

5.3 **Artigo 3:** Baixa ingestão de micronutrients está associada com eventos cardiovasculares em pacientes em atenção secundária**:** Uma análise transversal do estudo DICA Br.

Baixa ingestão de micronutrientes está associada com eventos cardiovasculares em pacientes em atenção secundária: Uma análise transversal do estudo DICA Br

Larissa Oliveira Chaves [a,*], Daniel Louzada Fernandes [b], Ana Luiza Gomes Domingos [a], Rodrigo Siqueira-Batista [c], Ângela Cristine Bersh-Ferreira [d], Aline Marcadenti [d, e], Bernardete Weber [d], Josefina Bressan[a]

[a]Department of Nutrition and Health, Universidade Federal de Viçosa, Viçosa – Minas Gerais, Brazil;

[b]Department of Informatics, Universidade Federal de Viçosa, Viçosa – Minas Gerais, Brazil;

[c]Department of Medicine and Nursing, Universidade Federal de Viçosa, Viçosa – Minas Gerais and Schoolof Medicine of the Faculdade Dinâmica do Vale do Piranga, Ponte Nova – Minas Gerais, Brazil;

[d]Research Institute, Hospital do Coração, São Paulo – São Paulo, Brazil;

[e]Graduate Program in Health Sciences (Cardiology), Institute of Cardiology/Fundação Universitária de Cardiologia do Rio Grande do Sul, Porto Alegre – Rio Grande do Sul, Brazil.

*Corresponding author: Larissa Oliveira Chaves, Department of Nutrition and Health, Universidade Federal de Viçosa, Avenida PH Rolfs s/n, Viçosa, Minas Gerais, CEP 36570-900, Brazil. Telephone: +5531-9-92862908. Email: larissa.chaves@ufv.br, larissaochaves@yahoo.com.br.

RESUMO

As doenças cardiovasculares (DCV) - conjunto de enfermidades relacionadas à alimentação não saudável - são as principais causas de morte em todo o mundo. Nesse cenário os micronutrientes se destacam por participar de processos anti-inflamatórios e antioxidantes relacionados à redução de fatores de riscos para o desenvolvimento de novos eventos cardiovasculares. O objetivo deste estudo foi avaliar se o número de eventos cardiovasculares é influenciado pela ingestão de micronutrientes, e se há diferenças metabólicas e de estilo de vida que impactam também na presença e em seu desenvolvimento. Este é um estudo transversal com dados basais do *"Brazilian Cardioprotective Nutritional Program - BALANCE Program*" que incluiu 1990 pacientes e suas características socioeconômicas, clínicas, bioquímicas e de estilo de vida. Os resultados demonstraram que os indivíduos com 2 ou mais eventos cardiovasculares, comparados aos indivíduos com apenas 1 evento, apresentavam menor renda, hábito de fumar, concentrações elevadas de glicemia e baixa concentração de lipoproteína de alta densidade, maior razão cintura estatura, maior consumo de alimentos culinários processados e menor consumo de fibras. Além disso, observou-se também a baixa ingestão de cobre, magnésio, selênio, vitaminas B12 e C. Em conclusão, os resultados sugerem relações importantes entre a baixa ingestão de micronutrientes e um perfil metabólico e alimentar adverso nos indivíduos com mais de um evento cardiovascular. Estudos com foco na ingestão de micronutrientes via suplementação e/ou alimentos fonte são encorajados, a fim de investigar os benefícios, os mecanismos envolvidos e as dosagens seguras para prevenção de um novo evento cardiovascular em populações cardiopatas.

PALAVRAS-CHAVES: Doenças cardiovasculares, micronutrientes, vitaminas, minerais, dieta, consumo alimentar.

INTRODUÇÃO

As doenças cardiovasculares (DCV) são a principal causa de morbidade e mortalidade em todo o mundo. Em 2019, 17,9 milhões de pessoas morreram devido a essas doenças, o que representou 32% de todas as mortes globais, produzindo um imenso impacto econômico e de saúde[1]. De acordo com a WHO (2019), as DCV ainda serão a principal causa de morte no mundo até 2030, aumentando para mais de 24 milhões por ano[2].

No entanto, é importante ressaltar que a maioria das DCV poderiam ser prevenidas, uma vez que alguns fatores de risco como o hábito de fumar, o sedentarismo e o consumo alimentar são modificáveis[3]. Nesse sentindo, destacamos então os micronutrientes, como as vitaminas e os minerais, sendo estes, componentes necessários para garantir uma saúde de qualidade. Os micronutrientes participam de processos metabólicos e antioxidantes, que tem papel fundamental no combate ao estresse oxidativo e nos danos mediados por radicais livres, promovendo um bom funcionamento do sistema nervoso e cardiovascular[5]. Para garantir um fornecimento adequado desses nutrientes, a dieta deve incluir grãos integrais, hortaliças e frutas, os quais são associados à diminuição da mortalidade por DCV[5]. Os dados na literatura indicam que a deficiência de micronutrientes aparece como problema de saúde global, atingindo cerca de 2 bilhões de pessoas no mundo[6].

Pesquisadores vêm demonstrando a importância dos micronutrientes na prevenção das DCV, como as vitaminas C e E, que estão associadas à melhora da aterosclerose[7,8] e a baixa ingestão das vitaminas B6 e B9 associadas ao Infarto Agudo do Miocárdio (IAM)[9]. Estudos clínicos apontam também associação entre a deficiência de micronutrientes e o aumento do risco para o desenvolvimento de DCV[10,11]. Entretanto, outros estudos não demonstraram evidências benéficas da suplementação de vitaminas em termos de redução do risco dessas doenças[12,13].

Por conseguinte, mais estudos clínicos com ênfase no papel dos micronutrientes na prevenção e no desenvolvimento das DCV são necessários para elucidar as relações metabólicas existentes. Nesse sentido, o objetivo deste estudo foi avaliar se o número de eventos cardiovasculares de pacientes em atenção secundária para DCV é influenciado pela ingestão de micronutrientes, e se há diferenças metabólicas e de estilo de vida que impactam também na presença e em seu desenvolvimento.

MATERIAIS E MÉTODOS

Participantes

Este é um estudo transversal com dados basais do estudo multicêntrico: "Brazilian Cardioprotective Nutritional Program (BALANCE Program)" registrado no ClinicalTrials.gov (NCT01620398). Fazem parte deste estudo 34 centros colaboradores das cinco regiões do Brasil.

Por se tratar de um estudo multicêntrico, cada Centro submeteu seu protocolo de estudo ao Comitê de Ética local, e os estudos foram iniciados após a aprovação de todos os protocolos[14]. Todos os pacientes incluídos no estudo assinaram um termo de consentimento livre e esclarecido. O protocolo do estudo foi desenvolvido de acordo com a Declaração de Helsinque e os princípios éticos brasileiros e internacionais[15].

Foram incluídos para este estudo 1990 participantes de ambos os sexos, com idade igual ou superior a 45 anos, com evidências atuais ou nos últimos dez anos de pelo menos um evento de DCV. Foram adotados os seguintes critérios para a confirmação, por meio de laudo médico, da DCV: (i) Doença Arterial Coronariana (DAC) presença de um ou mais sintomas: DAC assintomática, sintomática ou tratada e Infarto Agudo do Miocárdio (IAM); (ii) Doença Cerebrovascular: Acidente Vascular Cerebral (AVC), Ataque Isquêmico Transitório (AIT) e Acidente Vascular Encefálico (AVE); (iii) Doença Arterial Periférica (DAP) presença de um ou mais sintomas: DAP assintomática, sintomática ou tratada, amputação e aneurisma de aorta. Todos os critérios de elegibilidade são relatados no protocolo do estudo[14].

Avaliação sociodemográfica e de estilo de vida

Foram aplicados questionários próprios por entrevistadores treinados: (i) condições sociodemográficas: sexo, idade, renda familiar e nível de escolaridade; (ii) estilo de vida: atividade física e hábito de fumar; (iii) presença de doenças: Hipertensão Arterial Sistêmica (HAS), diabetes *mellitus*, dislipidemia e história familiar de DAC[14].

Avaliação antropométrica e de pressão arterial

O peso (kg), a altura (m) e o perímetro da cintura (PC) foram mensurados por pares de entrevistadores, utilizando a média de duas medidas. O PC foi avaliado a partir do ponto médio entre a borda inferior do arco costal e a crista ilíaca na linha axilar média[16]. O Índice de Massa Corporal (IMC) foi calculado pelo peso (kg)/altura (m)² para avaliar o estado

nutricional de adultos e idosos[17,18]. A relação cintura estatura (RCE) foi calculada pelo PC (cm) e altura (cm) como indicador de obesidade central e foi considerada alterada quando o resultado fosse maior igual 0,5[19].

A Pressão Arterial Sistólica (PAS) e a Pressão Arterial Diastólica (PAD) foram aferidas por profissionais de saúde treinados, utilizando a média de duas medidas, com esfigmomanômetro de mercúrio, seguindo as recomendações da *American Heart Association*[20].

Avaliações bioquímicas

Amostras de sangue foram coletadas após jejum de 12 a 14 horas. Marcadores clássicos de risco cardiometabólico como Triglicerídeos (TG), Colesterol Total (CT), glicemia de jejum e Lipoproteína de Alta Densidade (HDL) foram medidos pelo método enzimático colorimétrico (Johnsons & Johnsons, Raritan, EUA, VITROS 5600) e a Lipoproteína de Baixa Densidade (LDL) foi determinada pela equação de Friedewald[21]. A resistência à insulina foi estimada pelo índice de triglicerídeo-glicose (índice TyG), calculado pela fórmula abaixo (1)[22]. Para estimar disfunções de adiposidade visceral associadas ao risco cardiometabólico foi calculado o Índice de Adiposidade Visceral (IAV) pelas fórmulas abaixo para homens (2) e mulheres (3)[23].

$$Ln\left[\frac{TG\ em\ jejum\ (mg/dl) * glicemia\ em\ jejum\ (mg/dl)}{2}\right] (1)$$

$$\left[\frac{PC\ (cm)}{(39{,}69+1{,}88 * IMC(kg/m2))}\right] * \left(\frac{TG\ (mmol/L)}{1{,}03}\right) * \left(\frac{1{,}31}{HDL\ (mmol/L)}\right) (2)$$

$$\left[\frac{PC\ (cm)}{(36{,}58+1{,}89 * IMC\ (kg/m2))}\right] * \left(\frac{TG\ (mmol/L)}{0{,}81}\right) * \left(\frac{1{,}52}{HDL\ (mmol/L)}\right) (3)$$

Avaliação da ingestão alimentar

A ingestão alimentar foi avaliada pela média de dois recordatórios de 24 horas (R24H), e a ingestão de nutrientes foi estimada por meio do programa de computador Nutriquanti®[24]. Cada nutriente foi ajustado por 1000 kcal de ingestão de energia.

Com base nas informações fornecidas pelo R24H, os alimentos e as preparações também foram classificados pela NOVA. NOVA é uma classificação brasileira que reúne alimentos em quatro grupos (alimentos in natura ou minimamente processados; ingredientes culinários processados; alimentos processados; e alimentos ultra processados) de acordo com o processamento a que esses alimentos são submetidos[25]. As preparações mistas foram

classificadas de acordo com a proporção dos ingredientes principais. O perfil de ingestão alimentar foi expresso de acordo com a contribuição calórica dos alimentos agrupados em relação à ingestão energética diária ajustada para 1000 kcal de ingestão de energia.

Análise estatística

Os indivíduos foram classificados em dois grupos de acordo com o número de eventos cardiovasculares (1 evento ou ≥ 2 eventos). A normalidade das variáveis foi avaliada por meio do teste de Shapiro-Wilk, sendo constatado que os dados não seguem uma distribuição paramétrica. Dessa forma, os resultados são apresentados em mediana (intervalo interquartil). Para a comparação estatística entre os grupos foi realizado o teste não paramétrico de Mann–Whitney. O teste qui-quadrado de Pearson foi realizado para avaliar associações potenciais entre as características do estilo de vida (atividade física e hábito de fumar) e dados sociodemográficos.

As análises estatísticas foram realizadas utilizando a linguagem de programação Python (versão 3.7.6) por meio das bibliotecas numpy e scipy.stats para manipulação de dados e estatística, respectivamente.

RESULTADOS

O foco deste trabalho foi apresentar e elucidar diferenças metabólicas, de estilo de vida e ingestão alimentar entre os indivíduos que apresentam apenas 1 evento cardiovascular daqueles que apresentam 2 eventos ou mais. Dentre os eventos cardiovasculares apresentados por esta população observamos uma maior prevalência de DAC tratada (69.15%), seguido pelo IAM (49.35%), DAC sintomática (35.78%), DAC assintomática (16.48%), AVC (11.81%), DAP sintomática (6.83%), DAP assintomática (4.17%), DAP tratada (3.62%), aneurisma da aorta (1.96%) e amputação (1.56%) (dados não mostrados). Vale ressaltar que um mesmo indivíduo pode apresentar mais de um evento.

Dentre os 1990 participantes incluídos neste estudo 58.4% eram do sexo masculino e tinham mediana de idade de 63 anos (57-70). Os indivíduos com 2 eventos ou mais apresentaram menor renda, comparado ao grupo com apenas 1 evento. Além disso, observamos que os indivíduos dos dois grupos apresentaram outras doenças crônicas, como a obesidade, diabetes *mellitus*, dislipidemia, HAS e história de DAC nos dois grupos (Tabela 1).

Tabela 1: Caracterização geral dos indivíduos cardiopatas (n=1990) de acordo com o número de eventos cardiovasculares.

Variáveis	**1 evento (*n*=806)**	**≥ 2 eventos (*n*=1184)**	***p* valor**
Idade (anos)	63 (57-70)	62 (57-70)	0.169
Sexo			**0.006**
Feminino	365 (45.29)	462 (39.02)	
Masculino	441 (54.71)	722 (60.98)	
Nível de Escolaridade			0.181
Analfabeto/fundamental incompleto	244 (30.27)	314 (26.52)	
Fundamental I	252 (31.27)	400 (33.78)	
Fundamental II	105 (13.03)	157 (13.26)	
Ensino Médio	134 (16.63)	226 (19.09)	
Ensino Superior	71 (8.80)	87 (7.35)	
Renda Familiar			**0.020**
≤1 salários mínimos	105 (13.03)	164 (13.85)	
>1 e ≤3 salários mínimos	449 (55.70)	690 (58.28)	
≥4 e <10 salários mínimos	223 (27.67)	312 (26.35)	
≥10 salários mínimos	29 (3.60)	18 (1.52)	
Prevalência de doenças (sim)			
HAS	724 (89.83)	1063 (89.78)	0.966
Diabetes *mellitus*	345 (42.80)	512 (43.24)	0.882
Dislipidemia	634 (78.66)	934 (78.89)	0.948
Obesidade	314 (38.96)	439 (37.08)	0.423
História familiar de DAC	535 (66.38)	778 (65.71)	0.795

Dados apresentados em mediana (intervalo interquartil). Teste de Mann-Whitney para a comparação entre os grupos. Os dados apresentados por n (%), como frequências. Teste Qui-quadrado de Pearson. Valores estatisticamente significativos (p <0.05).

Abreviações: DAC – Doença Arterial Coronariana; HAS – Hipertensão Arterial Sistêmica. Salário mínimo (R$724.00 em 2014).

Em relação ao estilo de vida dos indivíduos observamos que a prática de atividade física não mostrou diferença entre os grupos, por outro lado, o hábito de fumar esteve mais presente nos indivíduos do grupo com 2 ou mais eventos (Tabela 2).

Tabela 2: Relação entre o número de eventos cardiovasculares e o estilo de vida dos indivíduos cardiopatas.

Variáveis	**1 evento (*n*=806)**	**≥ 2 eventos (*n*=1184)**	***p* valor**
Hábito de fumar			**0.007**
Sim	460 (57.07)	757 (63.94)	
Não	346 (42.93)	427 (36.06)	
Atividade Física			0.085
Sim	300 (37.22)	395 (33.36)	
Não	506 (62.78)	789 (66.64)	

Os dados estão apresentados por n (%), como frequências. Teste qui-quadrado de Pearson.

No que diz respeito às implicações metabólicas e clínicas, os resultados evidenciaram que os indivíduos do grupo com 2 ou mais eventos apresentaram elevadas concentrações de glicemia e menor concentração de HDL, assim como uma maior RCE, quando comparado ao grupo com apenas 1 evento (Tabela 3).

O consumo alimentar avaliado pelo grau de processamento dos alimentos de acordo com a NOVA mostrou que os indivíduos com 2 ou mais eventos apresentaram maior consumo de ingredientes culinários processados (p=0.004). Já os alimentos in natura/minimamente processados (p=0.097), alimentos processados (p=0.468) e alimentos ultra processados (p=0.120) não exibiram diferenças (dados não mostrados).

Tabela 3: Características metabólicas e clínicas dos indivíduos cardiopatas (n=1990) de acordo com o número de eventos cardiovasculares.

Variáveis	**1 evento (*n*=806)**	**≥ 2 eventos (*n*=1184)**	***p* valor**
Perímetro da cintura (cm)	100.00 (91.25-107.00)	100.00 (92.00-107.00)	0.294
IMC (kg/m^2)	28.57 (25.79-31.91)	28.38 (25.66-31.79)	0.230
RCE	0.61 (0.57-0.67)	0.62 (0.57-0.66)	**0.030**
IAV	2.31 (1.56-3.56)	2.40 (1.54-3.52)	0.252
PAD (mmHg)	80.00 (70.00-86.00)	80.00 (70.00-87.00)	0.457
PAS (mmHg)	130.00 (120.00-140.00)	130.00 (120.00-140.00)	0.112
Colesterol Total (mg/dl)	161.50 (138.00-193.75)	162.00 (138.00-193.00)	0.446
Triglicerídeos (mg/dl)	137.50 (99.00-190.00)	137.00 (99.00-190.00)	0.410
Glicemia (mg/dl)	102.00 (92.00-120.00)	103.00 (92.00-127.25)	**0.032**
HDL (mg/dl)	42.00 (35.00-50.00)	41.00 (34.00-49.00)	**0.011**
LDL (mg/dl)	86.50 (70.00-115.00)	88.50 (70.00-112.00)	0.384
Índice TyG	8.89 (8.51-9.29)	8.90 (8.53-9.35)	0.125

Dados apresentados em mediana (intervalo interquartil). Teste de Mann-Whitney para a comparação entre os grupos. Valores estatisticamente significativos (p <0.05).

Abreviações: HDL - *High Density Lipoprotein*; IAV – Índice de Adiposidade Visceral; IMC – índice de Massa Corporal; LDL –*Low Density Lipoprotein*; PAD – Pressão Arterial Diastólica; PAS – Pressão Arterial Sistólica; RCE – Razão Cintura Estatura; Índice TyG - Índice de Triglicerídeo-Glicose.

Os indivíduos com 2 ou mais eventos consumiram dietas mais ricas em lipídios e pobres em fibras. Ainda, foi observado uma menor ingestão de micronutrientes, sendo eles, magnésio, cobre, selênio, vitamina C e vitamina B12 (Tabela 4).

Tabela 4: Ingestão diária de nutrientes dos indivíduos cardiopatas de acordo com o número de eventos cardiovasculares.

Variáveis	**1 evento (*n*=806)**	**≥ 2 eventos (*n*=1184)**	***p* valor**
Calorias (kcal)	1331.93 (1051.16-1687.66)	1382.15(1079.55-1743.43)	0.056
Carboidratos (g)	176.52 (138.81-223.87)	181.52 (140.08-226.80)	0.118
Proteínas (g)	63.89 (46.74-85.53)	65.43 (48.42-88.49)	0.089
Lipídios (g)	38.97 (27.46-54.26)	40.87 (29.03-52.21)	**0.031**
AGS (g)	14.47 (11.74-17.22)	14.36 (11.43-17.36)	0.297
AGMI (g)	12.73 (10.48-15.17)	13.09 (10.36-15.73)	0.094
AGPI (g)	10.14 (8.23-12.39)	10.24 (8.37-10.24)	0.187
Trans (g)	0.03 (0.02-0.05)	0.03 (0.02-0.05)	0.344
Colesterol (mg)	178.72 (134.30-243.37)	181.03 (132.91-248.37)	0.313
Fibra alimentar (g)	19.15 (14.39-24.21)	17.82 (13.23-23.68)	**0.001**
Sódio (mg)	2696.18 (2320.65-3151.42)	2668.63 (2267.44-3110.00)	0.239
Cálcio (mg)	469.40 (323.34-665.79)	449.92 (314.12-635.74)	0.093
Ferro (mg)	6.27 (5.25-7.77)	6.26 (5.11-7.72)	0.306
Potássio (mg)	2014.25 (1719.84-2401.90)	2004.40 (1648.33-2363.22)	0.060
Magnésio (mg)	184.23 (155.60-217.73)	177.41 (148.88-215.21)	**0.003**
Fósforo (mg)	834.90 (720.79-992.86)	840.57 (716.12-993.82)	0.439
Cobre (mg)	0.91 (0.63-1.26)	0.84 (0.54-1.22)	**0.001**
Zinco (mg)	7.82 (5.98-10.43)	7.59 (6.01-10.34)	0.291
Selênio (µg)	19.78 (12.09-29.72)	18.11 (11.14-27.00)	**0.002**
Vitamina A (mg)	316.48 (176.82-517.23)	304.84 (165.59-505.62)	0.133
Vitamina B1 (mg)	0.72 (0.55-1.02)	0.70 (0.54-0.93)	0.069
Vitamina B2 (mg)	0.88 (0.64-1.21)	0.85 (0.59-1.21)	0.146
Vitamina B3 (mg)	12.04 (8.47-18.31)	12.57 (8.40-19.52)	0.170
Vitamina B5 (mg)	0.57 (0.24-1.08)	0.50 (0.24-0.99)	0.057
Vitamina B6 (mg)	0.48 (0.33-0.71)	0.48 (0.33-0.71)	0.477
Vitamina B7 (mg)	0.01 (0.00-0.04)	0.01 (0.00-0.04)	0.091
Vitamina B12 (mg)	0.68 (0.25-1.22)	0.59 (0.17-1.13)	**0.004**
Vitamina C (mg)	83.74 (34.70-181.22)	72.35 (30.49-157.42)	**0.003**
Vitamina D (µg)	4.52 (0.49-33.01)	4.17 (0.25-28.27)	0.075

Vitamina E (RE)	1.49 (1.11-2.21)	1.50 (1.09-2.18)	0.426
Ômega 3 (g)	0.96 (0.68-1.19)	0.94 (0.67-1.19)	0.128
Ômega 6 (g)	8.24 (6.51-10.34)	8.42 (6.62-10.37)	0.202

Dados apresentados em mediana (intervalo interquartil). Teste de Mann-Whitney para a comparação entre os grupos. Valores estatisticamente significativos ($p < 0.05$).

Legenda: Cada nutriente foi ajustado para 1.000 kcal de ingestão de energia.

Abreviações: AGMI – Ácido Graxo Monoinsaturado; AGPI – Ácido Graxo Poliinsaturado; AGS – Ácido Graxo Saturado; TRANS – Ácido Graxo *Trans*.

DISCUSSÃO

Implicações metabólicas, clínicas e de estilo de vida nas DCV

Em nosso estudo, observamos uma maior prevalência de DAC em relação aos outros eventos cardiovasculares. Essas doenças são de natureza multifatorial, causadas principalmente pela aterosclerose que é caracterizada pelo acúmulo de placas de colesterol nas paredes das artérias, o que causa obstrução do fluxo sanguíneo[26].

Sabe-se que a maioria das DCV podem ser prevenidas abordando os fatores de risco comportamentais, como, o hábito de fumar, que possui papel crucial na fisiopatologia das DCV[1]. O tabagismo é considerado um dos principais fatores de risco para DCV, uma vez que, a fumaça do tabaco contém partículas com efeitos pró-oxidantes que produzem radicais livres, aumentando assim, a peroxidação lipídica, desempenhando um papel na aterogênese, na oxidação de LDL e na redução de HDL[27]. Nesse sentido, no presente estudo o hábito de fumar esteve mais presente no grupo de indivíduos que apresentavam 2 ou mais eventos cardiovasculares.

Destaca-se também a relação do impacto da renda familiar no desenvolvimento das DCV. No presente trabalho, observamos que os indivíduos com renda mais baixa apresentavam mais de um evento cardiovascular. Pelo menos três quartos das mortes no mundo por DCV ocorrem em países de baixa e média renda, onde muitas das vezes não têm o benefício dos programas de atenção primária a saúde para a detecção e tratamento precoce[1]. Estudos apontam que pessoas com maior renda têm menor morbidade e mortalidade em quase todos os indicadores de saúde, longevidade e taxas de mortalidade[28-30]. Logo, infere-se que esta associação é devida ao maior acesso aos cuidados de saúde, melhor moradia e alimentação mais saudável[31].

Em relação ao perfil lipídico e glicemia, observamos que as concentrações de HDL se encontravam mais baixas e a glicemia mais elevada nos indivíduos com 2 ou mais eventos. Estudos epidemiológicos demonstram uma relação inversa entre baixa concentração de HDL e aumento do risco de DAC e IAM, e que o risco para DCV pode ser reduzido em indivíduos com altas concentrações de HDL[32]. Já as concentrações elevadas de glicose no sangue estão associadas a um maior risco de DCV e mortalidade, embora a magnitude do risco em diferentes populações seja menos clara[33].

Um marcador que tem demonstrado grande capacidade de identificar indivíduos em risco à saúde é a RCE, caracterizada por ser um índice antropométrico com poder de substituir o IMC e o PC por fornecer informações mais precisas de riscos cardiovasculares, que avalia a distribuição de gordura pelo corpo[19]. Além disso, a RCE está fortemente associada aos fatores de risco cardiovasculares e metabólicos, independentemente do peso corporal[29]. Uma revisão sistemática e metanálise realizada em 2012 concluiu que a RCE é a melhor ferramenta de triagem para detectar fatores de risco cardiometabólicos em ambos os sexos e diversos grupos étnicos, evidenciando sua superioridade sobre o IMC e o PC[34]**.** Esses dados corroboram com os resultados encontrados neste estudo, onde a RCE esteve mais elevada nos indivíduos com mais de um evento cardiovascular. Por isso, sugerimos que esta medida seja acrescentada na avaliação e acompanhamento dos pacientes em atenção primária e secundária.

Já se sabe que a alimentação tem um papel importante na prevenção e/ou no desenvolvimento das DCV. Estudos apontam que o risco de desenvolver DCV pode estar relacionado à elevada ingestão de lipídios na dieta, devido ao efeito pró-aterogênico, aumento da inflamação, sensibilidade à insulina e HAS[35,36]. Já a fibra alimentar, é frequentemente relatada como benéfica na redução do CT e da HAS, e, por isso, acredita-se que sua deficiência pode estar relacionada ao desenvolvimento de DCV[37]. Estudos clínicos que investigaram a ingestão de fibras sobre o risco dessas doenças relataram efeitos protetores[38,39]. Os achados na literatura podem explicar o elevado consumo de lipídios e baixo consumo de fibra nos indivíduos que apresentam 2 ou mais eventos cardiovasculares, comparados a aqueles com apenas 1 evento.

O papel dos micronutrientes nas DCV

Há evidências científicas de que as dietas enriquecidas com micronutrientes, incluindo, vitaminas e minerais, mantêm o status antioxidante celular, minimizando os impactos da atividade pró-inflamatória e do estresse oxidativo nos tecidos. Sabe-se que há

um aumento na produção de oxidantes com o avançar da idade, que pode ser neutralizado pelo aumento da ingestão de antioxidantes na dieta[40]. Desse modo, a ingestão adequada de micronutrientes, seja pela alimentação ou por meio da suplementação, pode oferecer principalmente aos idosos, proteção contra DCV[41]. No entanto, algumas revisões presentes na literatura não demonstraram evidências consistentes de que os suplementos têm efeitos sobre as DCV[42,43].

Observamos em nossos resultados uma menor ingestão de cobre (*Cu*), magnésio (*Mg*), selênio (*Se*), vitamina C e vitamina B12 nos indivíduos com 2 ou mais eventos cardiovasculares, comparados aqueles com apenas 1 evento, nos fazendo refletir sobre a importância da ingestão desses micronutrientes no estado de saúde de indivíduos cardiopatas.

Minerais

Os minerais têm importante função oxidante ou antioxidante que pode ter efeitos diretos na saúde cardiovascular, contribuindo principalmente no combate aos processos inflamatórios e de estresse oxidativo. No entanto, em desequilíbrio podem ser potencialmente perigosos e levar ao desenvolvimento de DCV, como DAC, cardiomiopatia, insuficiência cardíaca e arritmias[44].

Cobre

O *Cu* tem um papel importante tanto como um agente pró-oxidante quanto antioxidante, atuando principalmente como um cofator catalítico de enzimas, como, a superóxido dismutase (SOD), que é essencial na integridade do coração e dos vasos sanguíneos[45]. Logo, podemos concluir que um baixo consumo desse mineral pode comprometer o funcionamento de enzimas antioxidantes, como a SOD, podendo levar a um aumento do estresse oxidativo, e com isso aumentando o risco de desenvolvimento de novos eventos cardiovasculares. Esses dados corroboram com nossos resultados, uma vez que observamos um consumo de *Cu* significativamente menor nos indivíduos que apresentavam 2 ou mais eventos cardiovasculares, comparados aos com apenas 1 evento.

Os efeitos do *Cu* sobre o sistema cardiovascular ainda não estão consolidados. No entanto, já foram relatados benefícios [46], e observado que sua ingestão abaixo ou acima da RDA pode levar ao comprometimento da saúde cardiovascular [47]. De acordo com a RDA a recomendação de *Cu* é de 900 μg/dia em adultos e idosos e são encontrados em alimentos como fígado, amendoim, amêndoas, castanhas, mamão formosa, linhaça, gergelim, café,

entre outros [48]. Já em relação à suplementação de *Cu*, os efeitos benéficos ainda são controversos, mas alguns estudos apontam que a suplementação aumenta a atividade de SOD e reduz a hipertrofia cardíaca[49,50].

Os mecanismos pelos quais a deficiência de *Cu* pode prejudicar a saúde cardiovascular abrangem o aumento do estresse oxidativo devido à redução da atividade da SOD, e o acúmulo de glicose, fator central na fisiopatologia da aterosclerose. Além disso, pode causar HAS, arritmia cardíaca, aumento da inflamação e diminuição da coagulação do sangue[51]. Ainda assim, estudos clínicos devem ser encorajados, a fim de consolidar a relação existente entre a deficiência e/ou sobrecarga de *Cu* sobre o desenvolvimento das DCV e investigar as quantidades seguras e ideais de suplementação deste mineral.

Magnésio

Observamos em nossos resultados um menor consumo de *Mg* em indivíduos que apresentavam 2 ou mais eventos cardiovasculares comparados aos que apresentavam apenas 1 evento. Esses dados podem ser explicados uma vez que o *Mg* está envolvido em várias funções fisiológicas e bioquímicas, e que o aumento da ingestão alimentar e/ou suplementar desse mineral pode prevenir o estresse inflamatório e oxidativo, disfunção endotelial, aumento da reatividade do tônus vascular, melhora da pressão arterial, do perfil lipídico e resistência à insulina[52]. Estudos vêm apresentando benefícios da ingestão de *Mg* e/ou sua suplementação. O estudo de Posadas-Sanchez Et al (2016)[54] observaram baixas concentrações de *Mg* sérico associada negativamente a proteína C-reativa, conhecida por ser preditora da aterosclerose. Já o estudo de Qu et al (2013)[55] constataram que um aumento de 150 mg/dia a 400 mg/dia na ingestão de *Mg* reduziu 9% o risco de desenvolvimento de DCV.

Portanto, o consumo de alimentos ricos em *Mg*, como abacate, nozes, amêndoas, leguminosas, chocolate amargo, entre outros, é a melhor forma de manter as concentrações normais de *Mg*[52]. De acordo com a RDA, a quantidade recomendada de *Mg* em adultos e idosos é de 420 mg/dia e 320 mg/dia para homens e mulheres, respectivamente[53]. De acordo com os dados e estudos apresentados podemos concluir então que o baixo consumo de *Mg* pode desencadear desequilíbrios metabólicos importantes e pode estar diretamente relacionado ao desenvolvimento de eventos cardiovasculares.

Selênio

Nossos resultados apontaram um menor consumo de *Se* em indivíduos pertencentes ao grupo que apresentavam 2 ou mais eventos cardiovasculares comparados aos com apenas 1 evento. Sabe-se que este mineral possui poder antioxidante prevenindo o estresse oxidativo e a inflamação, além de inibir a toxicidade de metais pesados como o mercúrio, prata e cádmio, que é um fator de risco para aterosclerose[56]. Portanto, entende-se que a ingestão adequada de *Se* pode prevenir o desenvolvimento de eventos cardiovasculares, ao passo que o contrário pode acarretar em seu desenvolvimento, o que justifica os resultados encontrados em nosso trabalho. É importante destacar então a recomendação de ingestão de *Se,* que de acordo com a RDA para homens e mulheres é de 55mg/dia, e as principais fontes deste mineral são os peixes, carnes de boi, frango, queijos, leite desnatado, ovos e castanhas[57].

Um estudo mostrou efeitos positivos do *Se* na redução do estresse oxidativo, LDL, glicose e a resistência à insulina, e consequentemente observaram um aumento nas concentrações plasmáticas de glutationa e óxido nítrico[58]. Outro estudo verificou ainda que a deficiência de Se foi associada com o aumento da mortalidade cardiovascular[59]. Em uma metanálise, os pesquisadores observaram que as concentrações plasmáticas de *Se* tiveram uma associação inversa com o risco de DCV, no entanto, a suplementação não teve efeitos significativos sobre os eventos cardiovasculares[60]. Essas descobertas inconsistentes sobre a associação entre as concentrações de *Se* e eventos cardiovasculares podem ser devido a alguns fatores como à absorção, biodisponibilidade, variações metodológicas, tamanho da amostra, duração do estudo, etnia dos participantes e também a dosagem de suplemento[60].

Vitaminas

As vitaminas antioxidantes (B6, B9, B12, C e E) têm importante papel na prevenção e na terapia das DCV, pois, são capazes de melhorar o sistema cardiovascular inibindo o estresse oxidativo, inflamação e a disfunção endotelial[61]. Estudos clínicos, *in vitro* e experimentais, sugerem efeitos benéficos das vitaminas na inibição da aterogênese e progressão da aterosclerose[62]. Além disso, estudos de coorte e clínicos mostraram benefício da suplementação de vitaminas. No entanto, um estudo de revisão observou que a maioria dos ensaios clínicos randomizados não puderam evidenciar a eficácia das vitaminas nas DCV[62].

Vitamina B12

A vitamina B12, também conhecida como cobalamina, está presente em alimentos de origem animal, como a carne, os ovos e o leite. Contudo, é importante ressaltar que a capacidade de absorção se altera com a idade, afetando sua biodisponibilidade[63]. De acordo com a RDA a quantidade adequada de ingestão é de 2,4μg/dia para adultos e idosos[64].

Em nosso estudo, observamos um menor consumo de vitamina B12 em indivíduos que apresentavam 2 ou mais eventos cardiovasculares comparados aos com apenas 1 evento. Um fator importante e que pode explicar esse achado é que essa vitamina é fundamental no metabolismo da homocisteína total (Hcy_t), uma vez que sua deficiência pode aumentar a concentração de Hcy_t, e ela tem sido implicada no desenvolvimento de DCV[65]. Essa associação é apoiada por mecanismos fisiopatológicos, como função endotelial prejudicada, aumento do estresse oxidativo, oxidação da LDL, redução da atividade antioxidante e da síntese de óxido nítrico, que tem efeitos vaso- anticonstritores e vaso-antiagregantes, inflamação e rigidez arterial[65].

Estudos realizados em humanos têm demonstrado efeitos benéficos da ingestão adequada e dos níveis tissulares de vitamina B12 e DCV. Um estudo com indivíduos chineses mostrou concentrações elevadas de vitamina B12 e baixa de Hcy_t em pacientes que apresentavam DCV[66]. Evidências também mostraram que concentrações elevadas de Hcy_t e baixas de vitamina B12 em mulheres, têm forte associação com o maior risco de todas as causas e mortes por DCV na população idosa[67].

De acordo com os dados apresentados, notamos que há evidências de que a vitamina B12 tem forte impacto na redução da Hcyt, desempenhando um papel na prevenção de DCV. No entanto, mais estudos clínicos devem ser realizados, a fim de compreendermos melhor os mecanismos associados.

Vitamina C

A vitamina C, conhecida como ácido ascórbico, tem função antioxidante protegendo contra o estresse oxidativo, e pode ser encontrada no brócolis, couve, caju, goiaba, acerola, pimentão amarelo, entre outros[68]. De acordo com a RDA, a ingestão diária recomendada é de 75 mg para mulheres e 90 mg para homens[64].

Sabe-se que as funções da vitamina C como antioxidante e como cofator enzimático estão bem estabelecidas, mas, os mecanismos envolvidos nas DCV não estão totalmente elucidadas[68]. No ano de 1996, o estudo de Weber, Erl & Weber já havia demonstrado o

benefício da vitamina C em reduzir a adesão de monócitos ao endotélio. Essa adesão é a chave na formação de ateromas e é considerado um dos primeiros sinais do desenvolvimento da aterosclerose[69]. Além disso, a vitamina C demonstrou melhorar a produção de óxido nítrico, que, por sua vez, aumenta a vasodilatação, reduzindo a pressão arterial[70]. Em 2008, uma metanálise baseada em 14 estudos, com mediana de acompanhamento de 10 anos, concluiu que a vitamina C na dieta tem uma associação inversa com o risco de DCV, ao passo que a ingestão de suplementos não teve associação significativa com o risco de DCV[71]. Estes dados corroboram com os nossos achados, uma vez que revelamos um menor consumo de vitamina C em indivíduos com 2 ou mais eventos cardiovasculares, onde apresentam um perfil metabólico mais comprometido, comparados aos com apenas 1 evento.

Portanto, a vitamina C tem um forte potencial para reduzir o risco cardiovascular, mas os resultados são inconclusivos principalmente em relação a sua suplementação. Por isso, mais estudos devem ser realizados a fim de entendermos melhor a relação dessa vitamina com o desenvolvimento de eventos cardiovasculares e seus potenciais mecanismos.

Limitações e pontos fortes

Este estudo apresenta algumas limitações, principalmente a natureza transversal o que não possibilita o estabelecimento de relações de causa-efeito, permitindo apenas associações. No entanto, tem significância biológica e clínica, uma vez que está relacionado à DCV, sendo importante para orientar e sugerir novos estudos de intervenção ou de caso-controle que poderia determinar possíveis mecanismos envolvidos neste campo.

O ponto mais forte deste trabalho é que se trata de um estudo multicêntrico, que avaliou um grande número de participantes com DCV em todas as regiões do Brasil, possibilitando a análise de uma variação maior de dados. Além disso, a qualidade dos dados foi maximizada por meio da automação da entrada de dados, do contato mensal com os pesquisadores e do monitoramento das visitas aos centros.

CONCLUSÕES

Conclui-se que os resultados alcançados sugerem relações importantes entre a baixa ingestão de algumas vitaminas (B12 e C) e minerais (Cu, Mg e Se) e a presença de mais de um evento cardiovascular em indivíduos cardiopatas, ratificando a presença de um perfil metabólico favorável para o aparecimento das DCV. Como demonstrado, tal fato pode estar relacionado principalmente a fatores inflamatórios e de estresse oxidativo, entre outros, que podem desencadear risco aumentado para as DCV. Esse é um ponto de partida interessante

para o desenvolvimento de estudos clínicos com foco na ingestão de micronutrientes, via suplementação e/ou alimentos fonte, a fim de investigar as dosagens, uma vez que os estudos existentes não são claros quanto à quantidade adequada e segura para garantir os benefícios destes micronutrientes, os reais benefícios e os mecanismos envolvidos para prevenção de um novo evento cardiovascular em populações cardiopatas,

Conflict of interest

The authors declare no conflict interest.

Statement of authorship

The conception and design of the study were performed by ACBF, AM and BW. The generation and data collection were performed by LOC, ACBF, BW and JB. The assembly and analysis and/or interpretation of the data were performed by LOC, DLF, ALGD, RSB and JB. All authors contributed in drafting the manuscript, final approval for publishing submission, and ensuring the integrity and accuracy of the work.

Funding sources

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Acknowledgements

We thank all the patients for participating in this project, and all participating centers: Cristiane Kovacs - Instituto Dante Pazzanese de Cardiologia, São Paulo-SP. Annie S B Moreira - Hospital Universitário Pedro Ernesto, Rio de Janeiro-RJ e Instituto Nacional de Cardiologia, Rio de Janeiro-RJ. Rosileide S Torres - Hospital das Clínicas Gaspar Vianna, Belém-PA. Helyde A Marinho - Instituto Nacional de Pesquisas da Amazônia, Manaus-AM. Cristina H de Matos - Universidade Vale do Itajaí, Itajaí-SC. Renata T A Bertacco - Universidade Federal de Pelotas, Pelotas-RS. Gabriela C Souza - Hospital de Clínicas de Porto Alegre, Porto Alegre-RS. Gabriela S Shirmann - Universidade da Região da Campanha, Bagé-RS. Francisca E Z Nagano - Hospital de Clínicas da Universidade Federal do Paraná, Curitiba-PR. Maria E M Ramos - Hospital Universitário Associação Educadora São Carlos, Canoas-RS. Soraia Poloni - Instituto de Cardiologia do Rio Grande do Sul, Porto Alegre-RS. Raquel M El Kik - Hospital São Lucas da Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre-RS. Naoel H Feres - Universidade Federal do Mato Grosso, Cuiabá-MT. Eliane S Dutra - Hospital Universitário de Brasília, Brasília-DF. Ana P

P F Carvalho - Hospital das Clínicas de Goiânia, Goiânia-GO. Marta M David - Hospital Universitário Maria Aparecida Pedrossian, Campo Grande-MS. Isa G Rodrigues - Pronto Socorro Cardiológico Universitário de Pernambuco, Recife-PE. Antonio C S Sousa - Hospital São Lucas, Aracaju-SE. Amanda G L Coura - Hospital Universitário Alcides Carneiro, Campina Grande-PB. Josilene M F Pinheiro - Hospital Universitário Ana Bezerra, Santa Cruz-RN. Sandra M L Vasconcelos - Universidade Federal de Alagoas, Maceió-AL. Andreza M Penafort - Universidade de Fortaleza, Fortaleza-CE. Daniele M O Carlos - Hospital de Messejana, Fortaleza-CE. Viviane Sahade - Hospital Universitário Professor Edgard Santos, Salvador-BA. Adriana B Luna - Hospital Universitário da Universidade Federal de Sergipe, Aracaju-SE. José A F Neto - Hospital Universitário da Universidade Federal do Maranhão, São Luís-MA. Emilio H Moriguchi - Associação Veranense de Assistência em Saúde, Veranópolis-RS. Maria C O Izar - Universidade Federal de São Paulo, São Paulo-SP. Sônia L Pinto - Universidade Federal de Tocantins, Palmas-TO. Hospital São Vicente de Paulo, Luciano M Backes - Passo Fundo-RS. Simone R Souza - Instituto Estadual de Cardiologia Aloysio de Castro, Rio de Janeiro-RJ. Magali C C - COTENUT, Porto Alegre – RS.

REFERÊNCIAS

1. World Health Organization (WHO). Cardiovascular diseases (CVDs), 2021. Disponível em <http://origin.who.int/cardiovascular_diseases/en/>.

2. World Heart Federation, 2019. https://www.world-heart federation.org/world-heart-day/world-heart-day-2019/. Acessado em 9 de Julho de 2021.

3. BRASIL. VIGITEL, BRASIL 2019: vigilância de fatores de risco e proteção para doenças crônicas por inquérito telefônico. Brasília: Ministério da Saúde, 2020.

4. Moser MA, Chun OK. Vitamin C and heart health: A review based on findings from epidemiologic studies. Int J Mol Sci,17(8), 2016.

5. Ilow R, Regulska-Ilow B, Różańska D, et al. Evaluation of mineral and vitamin intake in the diet of a sample of Polish population: baseline assessment from the prospective cohort 'PONS' study. Ann Agric Environ Med, 18(2):235-40, 2011.

6. Leão ALM, Santos LC. Consumo de micronutrientes e excesso de peso: existe relação? RevBras Epidemiol, 15(1):85-95, 2012.

7. Mathur P, Ding Z, Saldeen T, et al. Tocopherols in the prevention and treatment of atherosclerosis and related cardiovascular disease. Clin Cardiol, 38(9):570–576, 2015.

8. Moser MA, Chun OK. Vitamin C and heart health: A review based on findings from epidemiologic studies. Int J Mol Sci, 17(8), 2016.

9. Rimm EB, Willett WC, Hu FB, et al. Folate and vitamin B6 from diet and supplements in relation to risk of coronary heart disease among women. JAMA, 279(5):359–364, 1998.

10. Toole TF, Mahnow MR, Chambless LE, Spence, et al. Lowering homocysteine in patients with ischemic strole to prevent recurrent stroke, myocardial infarction and death: the Vitamin Intervention for Stroke Prevention (VISP) randomized controlled-trial. JAMA, 291:565-75, 2004.

11. Spence JD, Bang H, Chambles LE, et al. Vitamin Intervention for Stroke Prevention trial: an efficacy analysis.Stroke, 36(11):2404-9, 2005.

12. Leone N, Courbon D, Ducimetiere P, et al. Zinc, copper and magnesium and risks for all cause, cancer, and cardiovascular mortality. Epidemiology, 17(3): 308-14, 2006.

13. Ebbing M, Bleie O, Ueland PM, et al. Mortality and cardiovascular events in patients treated with homocysteine-lowering B vitamins after coronary angiography: a randomized controlled trial. JAMA, 300(7):795-804, 2008.

14. Weber B, Bersch-Ferreira AC, Torreglosa CR, et al. The Brazilian Cardioprotective Nutritional Program to reduce events and risk factors in secondary prevention for cardiovascular disease: study protocol (The BALANCE Program Trial). American Heart Journal,171:73–81, 2016.

15. World Medical Association. World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects. Journal of the American Medical Association,310:2191–4, 2013.

16. World Health Organization (WHO). Waist circumference and waist–hip ratio: report of a WHO expert consultation. *WHO*, 2008.

17. World Health Organization (WHO). Physical status: the use an interpretation of anthropometry. Technical Report Series. Geneva: *WHO*, 1995.

18. OPAS – Organização Pan-Americana da Saúde. XXXVI Reunión Del Comitê Asesor de Investigaciones em Salud – Encuestra Multicêntrica – Salud Beinestary Envejecimeiento (SABE) en América Latina e el Caribe - Informe preliminar. Disponívelem<URL: //WWW.opas.org/program/sabe.htm.> 2002.

19. Ashwell M, Gunn P, Gibson S. Waist-to-height ratio is a better screening tool than waist circumference and BMI for adult cardiometabolic risk factors: systematic review and meta-analysis. Obesity Reviews, 13:275-278, 2012.

20. AHA. Understanding blood pressure Reading site. 2018.

21. Friedewald WT, Levy RI, Fredrickson DS. Estimation of the concentration of low-density lipoprotein cholesterol in plasma, without use of the preparative ultracentrifuge. Clin Chem, 18(6):499–502, 1972.

22. Simental-mendía LE, Rodríguez-morán M, Guerrero-romero F. The product of fasting glucose and triglycerides as surrogate for identifying insulin resistance in apparently healthy subjects. Metabolic Syndrome and Related Disorders, 6:299–304, 2008.

23. Amato MC, Giordano C, Galia M, et al. Visceral Adiposity Index: a reliable indicator of visceral fat function associated with cardiometabolic risk. Diabetes Care, 33:920-2, 2010.

24. Galante AP. Desenvolvimento e validação de um método computadorizado para avaliação do consumo alimentar, preenchido por indivíduos adultos utilizando a Web. Tese (Doutorado em Nutrição Humana Aplicada) – Faculdade de Ciências Farmacêuticas, Universidade de São Paulo, 2007.

25. Monteiro CA, Cannon G, Levy R, et al. O Sistema Alimentar. World,7:1–3, 2016.

26. Center for disease control and prevention, 2021.

27. Bagchi D, Sen CK, Ray SD.et al. Molecular mechanisms of cardioprotection by a novel grape seed proanthocyanidin extract. Mutat Res, 523–524, 87–97, 2003.

28. Braveman PA, Cubbin C, Egerter S, et al. Socioeconomic disparities in health in the United States: what the patterns tell us. Am J Public Health, 100(suppl 1):S186-S196, 2010.

29. Galea S, Tracy M, Hoggatt KJ, et al. Estimated deaths attributable to social factors in the United States. Am J Public Health, 101(8):1456-1465, 2011.

30. Chetty R, Stepner M, Abraham S, et al. The association between income and life expectancy in the United States, 2001-2014. JAMA, 315(16):1750-1766, 2016.

31. Abdall SM, Galea SS. Trends in Cardiovascular Disease Prevalence by Income Level in the United States. JAMA Network Open, 3(9):e2018150, 2020.

32. Castelli WP. Cholesterol and lipids in the risk of coronary artery disease—the Framingham Heart Study. Can J Cardiol,4 Suppl A:5A–10A, 1988.

33. Di Angelantonio E, Sarwar N, Perry P, et al. Major lipids, apolipoproteins, and risk of vascular disease. JAMA, 302: 1993–2000, 2009.

34. Ashwell M, Gunn P, Gibson S. Waist-to-height ratio is a better screening tool than waist circumference and BMI for adult cardiometabolic risk factors: systematic review and meta-analysis. Obes Rev, 13(3):275-86, 2012.

35. Niinikoski H, Jula A, Viikari J, et al. Blood pressure is lower in children and adolescents with a low-saturated-fat diet since infancy: the special turku coronary risk factor intervention project. Hypertension, 53:918–24, 2009.

36. Morandi A, Fornari E, Opri F, et al. High-fat meal, systemic inflammation and glucose homeostasis in obese children and adolescents. Int J Obes, 41:986–9, 2017.
37. Sánchez-Muniz FJ. Dietary fibre and cardiovascular health. Nutr Hosp, 27(1):31–45, 2012.
38. Kokubo Y, Iso H, Saito I. Dietary fiber intake and risk of cardiovascular disease in the Japanese population: the Japan Public Health Center-based study cohort. Eur J Clin Nutr, 65(11):1233–1241, 2011.
39. Buil-Cosiales P, Zazpe I, Toledo E. Fiber intake and all-cause mortality in the Prevención con Dieta Mediterránea (PREDIMED) study. Am J Clin Nutr, 100(6):1498–1507, 2014.
40. Hagen TM, Liu J, Lykkesfeldt J, et al. Feeding acetyl-l-carnitine and lipoic acid to old rats significantly improves metabolic function while decreasing oxidative stress. Proc Natl Acad Sci USA, 99:1870–5, 2002.
41. Visioli F, Hagen TM. Nutritional strategies for healthy cardiovascular aging: Focus on micronutrients. Pharmacological Research, 55:199–206, 2007.
42. Fortmann SP, Burda BU, Senger CA, et al. Vitamin and mineral supplements in the primary prevention of cardiovascular disease and cancer: An updated systematic evidence review for the U.S. Preventive Services Task Force. Ann Intern Med, 159:824–34, 2013.
43. Moyer VA, Force USPST. Vitamin, mineral, and multivitamin supplements for the primary prevention of cardiovascular disease and cancer: U.S. Preventive services Task Force recommendation statement. Annals of Internal Medicine, 160:558–64, 2014.
44. Kolte DK, Vijayaraghavan S, Khera DA, et al. Role of magnesium in cardiovascular diseases. Cardiology in Review, 22:182–92, 2014.
45. Al-Bayati MA, Jamil DA, Al-Aubaidy, HA. Cardiovascular effects of copper deficiency on activity of superoxide dismutase in diabetic nephropathy. North American Journalof Medical Sciences, 7:41–46, 2015.
46. Kang YJ. Copper and homocysteine in cardiovascular diseases. Pharmacology & Amp; Therapeutics, 129:321–31, 2011.
47. Medeiros DM. Perspectives on the Role and Relevance of Copper in Cardiac Disease. Biological Trace Element Research, 176:10–19 2017.
48. Institute of Medicine. Dietary reference intakes for vitamin A, vitamin K, arsenic, boron, chromium, copper, iodine, iron, manganese, molybdenum, nickel, silicon, vanadium, and zinc. Washington (DC): National Academy Press; 2002.

49. Zhou Z, Johnson WT, Kang YJ. Regression of copper-deficient heart hypertrophy: reduction in the size of hypertrophic cardiomyocytes. Journal of Nutritional Biochemistry, 20:621–8, 2009.

50. Duncan C, White AR. Copper complexes as therapeutic agents. Metallomics, 4:127–38, 2012.

51. Bost M, HoudartS, Oberli M, et al. Dietary copper and human health: Current evidence and unresolved issues. Journal of Trace Elements in Medicine and Biology, 35:107–15, 2016.

52. Fang X, WangK, Han D, et al. Dietary magnesium intake and the risk of cardiovascular disease, type 2 diabetes, and all-cause mortality: a dose–response meta-analysis of prospective cohort studies. BMC Medicine, 14:210, 2016.

53. Posadas-Sanchez R, Posadas-Romero C, Cardoso-Saldana G, et al.Serum magnesium is inversely associated with coronary artery calcification in the Genetics of Atherosclerotic Disease (GEA) study. Nutrition Journal, 15:22, 2016.

54. Qu X, Jin F, Hao Y, et al. Magnesium and the risk of cardiovascular events: a meta-analysis of prospective cohort studies. PLoS One, 8:e57720, 2013.

55. Institute of Medicine. Dietary reference intakes for calcium, phosphorus, magnesium, vitamin D, and fluoride. Washington (DC): National Academy Press; 1997.

56. Hu XF, Eccles KM, Chan HM. High selenium exposure lowers the odds ratios for hypertension, stroke, and myocardial infarction associated with mercury exposure among Inuit in Canada. Environment International, 102:200–06, 2017.

57. Institute of Medicine. Dietary reference intakes for vitamin C, vitamin E, selenium, and carotenoids. Washington (DC): National Academy Press; 2000.

58. Panchal SK, WanyonyiS, Brown L. Selenium, Vanadium, and Chromium as Micronutrients to Improve Metabolic Syndrome. Current Hypertension Reports, 19:10, 2017.

59. Alehagen U, JohanssonP, BjornstedtM, et al. Relatively high mortality risk in elderly Swedish subjects with low se, et al. Selenium status. European Journal of Clinical Nutrition, 70:91–96, 2016.

60. Zhang X, Liu C, Guo J. Selenium status and cardiovascular diseases: meta-analysis of prospective observational studies and randomized controlled trials. European Journal of Clinical Nutrition, 70:162–9, 2016.

61. Lugrin J, Rosenblatt-Velin N, Parapanov R, et al. The role of oxidative stress during inflammatory processes. Biol Chem, 395(2):203–230, 2014.

62. Balazs D, Laszlo D. Role of vitamins in cardiovascular health and disease. Research Reports in Clinical Cardiology, 5 283–295, 2014.

63. Roman P. Is Vitamin B12 Deficiency a Risk Factor for Cardiovascular Disease in Vegetarians?.Am J Prev Med, 48(6):e11–e26, 2015.

64. Padovani RM, Amaya-Farfán J, Colugnati FAB, et al. Dietary reference intakes: application of tables in nutritional studies. Rev Nutr, 19(6), 2006.

65. Yuan AM,CarterMP, Burgess S et al. Homocysteine, B vitamins, and cardiovascular disease: a Mendelian randomization study Shuai. BMC Medicine, 19:97, 2021.

66. Mao X, Xing X, Xu R, et al. Folic acid and vitamins D and B12 correlate with homocysteine in Chinese patients with type-2 diabetes mellitus, hypertension, or cardiovascular disease. Medicine, 95:e2652, 2016.

67. Mendonça N, Jagger C, Granic A, et al. Elevated total homocysteine in all participants and plasma vitamin B12 concentrations in women are associated with all-cause and cardiovascular mortality in the very old: The Newcastle 85+ Study. J. Gerontol. Ser, 73:1258–1264, 2018.

68. Morelli MB, Gambardella J, Castellanos V, et al. Vitamin C and Cardiovascular Disease: An Update. Antioxidants, 9:1227, 2020.

69. Weber C, Erl W, Weber K, et al. Increased Adhesiveness of Isolated Monocytes to Endothelium Is Prevented by Vitamin C Intake in Smokers. Circulation, 93:1488–1492, 1996.

70. D'uscio LV, Milstien S, Richardson D, et al. Long-term vitamin C treatment increases vascular tetrahydrobiopterin levels and nitric oxide synthase activity. Circ. Res, 92:88–95, 2003.

71. Ye Z, Song H. Antioxidant vitamins intake and the risk of coronary heart disease: Meta-analysis of cohort studies. Eur J Cardiovasc Prev Rehabil, 15:26–34, 2008.

6 CONCLUSÕES GERAIS

A revisão sistemática resumiu as últimas informações sobre a utilização de algoritmos ML para avaliar o consumo alimentar, e pode servir como um guia para profissionais da nutrição que queiram trabalhar com a IA. Atualmente há um grande e crescente interesse na utilização de Algoritmos ML na área da nutrição, principalmente devido a um significativo aumento das publicações nos últimos anos. Os algoritmos de aprendizagem supervisionada foram os mais utilizados, assim como o questionário de frequência alimentar para avaliar o consumo alimentar. Mesmo compreendendo a importância de investigar o consumo alimentar em diferentes populações, e como o uso de algoritmos ML pode ser interessante, houve pouca diversidade de países envolvidos nos estudos analisados. Encoraja-se a utilização desses algoritmos na investigação do consumo alimentar em diferentes populações, uma vez que os problemas enfrentados são distintos e requerem diferentes investigações e intervenção para o desenvolvimento de programas de reeducação alimentar e políticas públicas específicas.

Os resultados revelaram diferenças relacionadas com o sexo e gênero na DCV que poderiam levar a diferentes eventos cardiovasculares. Assim, são necessários estudos centrados nessas diferenças na DCV, incluindo ensaios clínicos para elucidar os mecanismos envolvidos. Constatamos que a aplicação dos algoritmos de ML não supervisionados identificaram quatro grupos de perfil distintos, que apresentavam diferenças significativas, de uma forma altamente precisa. Assim, demonstra-se mais uma vez que as técnicas ML constituem uma ferramenta inovadora na área da saúde, especialmente para identificar padrões de forma rápida e eficaz, e apontar alternativas para futuras investigações.

Além disso, os resultados também sugerem relações importantes entre baixa renda, hábito de fumar, alta RCE e concentração de glicemia e baixa concentração de HDL, além de alta ingestão de lipídios e baixa de fibras, vitaminas B12 e C e dos minerais Cu, Mg e Se com a presença de mais de um evento cardiovascular em indivíduos cardiopatas. Esse é um ponto de partida interessante para o desenvolvimento de estudos clínicos com foco na ingestão de micronutrientes, via suplementação e/ou alimentos fonte, a fim de investigar seus reais benefícios e os mecanismos envolvidos.

7 REFERÊNCIAS

ABEP. **Critérios de Classificação Econômica do Brasil**, 2013. Disponível em: <http://www.abep.org/>.

AHA. **Understanding blood pressure Reading**, 2018.

AHMAD, M. A.; ECKERT, C.; TEREDESAI, A. Interpretable machine learning in healthcare. In Proceedings of the 2018 ACM International Conference on Bioinformatics. **Computational Biology, and Health Informatics**. p. 559–560, 2018.

AL'AREF, S. J. MALIAKAL, G.; SINGH, G.; ROSENDAEL, A. R. V.; MA, X.; XU, Z. et al. Machine learning of clinical variables and coronary artery calcium scoring for the prediction of obstructive coronary artery disease on coronary computed tomography angiography: analysis from the CONFIRM registry. **European Heart Journal**, v. 41, n. 3, p. 1–9. 2019.

AL-MAQALEH, B. M.; ABDULLAH, A. M. G. Intelligent predictive system using classification techniques for heart disease diagnosis. **International Journal of Computer Science Engineering (IJCSE)**, v. 6, n. 6, 145-151, 2017.

AMATO, M. C.; GIORDANO, C.; GALIA, M.; CRISCIMANNA, A.; VITABILE, S.; MIDIRI, M. et al. Visceral Adiposity Index: a reliable indicator of visceral fat function associated with cardiometabolic risk. **Diabetes Care**, v. 33, n. 4, p. 920-2, 2010.

ASHWELL, M.; GUNN, P.; GIBSON, S. Waist-to-height ratio is a better screening tool than waist circumference and BMI for adult cardiometabolic risk factors: systematic review and meta-analysis. **Obesity Reviews**, v. 13, p. 275-278, 2012.

ATUN, R. Transitioning health systems for multimorbidity. **Lancet**, v. 386, p. 721-2, 2015.

BABU, S.; FAMINA, K. P.; VIVEK, E. M.; FIDA, K. Heart disease diagnosis using data mining technique. In Electronics, Communication and Aerospace Technology (ICECA), International conference, v. 1, p. 750-753, 2017.

BASTIDE, N. M.; PIERRE, F. H. F.; CORPET, D. E. Heme iron from meat and risk of colorectal cancer: a meta-analysis and a review of the mechanisms involved. **Cancer Prevention Research**, v. 4, n. 2, p. 177-184, 2011.

BEN-ISRAEL, D.; JACOBS, W. B.; CASHA, S; LANG, S.; RYU, W. H. A.; LOTBINIERE-BASSETT, M. et al. The impact of machine learning on patient care: A systematic review. **Artificial Intelligence In Medicine**, v. 103, 2020.

BRASIL. Ministério da Saúde. Secretaria de Vigilância em Saúde. **Plano de ações estratégicas para o enfrentamento das doenças crônicas não transmissíveis (DCNT) no Brasil 2011-2022 / Ministério da Saúde**. Secretaria de Vigilância em Saúde. Departamento de Análise de Situação de Saúde. – Brasília : Ministério da Saúde. 160 p., 2011a.

BRASIL. Pesquisa de Orçamentos Familiares 2008-2009. **Tabelas de Composição Nutricional dos Alimentos Consumidos no Brasil**. Rio De Janeiro, 2011b.

BRASIL. **VIGITEL, BRASIL 2019: vigilância de fatores de risco e proteção para doenças crônicas por inquérito telefônico**. Brasília: Ministério da Saúde, 2020.

BRUM, F.; MOZZAQUATRO, P. M.; ZANATTA, J. M. Estudo sobre os algoritmos de clusterização *Hierarchical Clusteres* e *Simple K-means* aplicados no agrupamento de padrões similares. **Revista da Universidade Vale do Rio Verde**, v. 17, n. 1, 2019.

CARVALHO, D. R.; MOSER, A. D.; DA SILVA, V. A.; DALLAGASSA, M. R. Mineração de dados aplicada à fisioterapia. **Fisioterapia em Movimento**, v. 25, n. 3, p. 595-605, 2012.

CERQUEIRA, F. R. FERREIRA, T. G.; OLIVEIRA, A. P.; AUGUSTO, D. A.; KREMPSER, E.; BARBOSA, H. J. C. et al. Nicesim: an open-source simulator based on machine learning techniques to support medical research on prenatal and perinatal care decision making. **Artificial intelligence in medicine**, v. 62, n. 3, p. 193–201, 2014.

CHADHA, R.; SHUBHANKAR, M.; VARDHAN, A.; PRADHAN, T. Application of data mining techniques on heart disease prediction: a survey. **Emerging Research**

in Computing, Information, Communication and Applications. New Delhi, p. 413-426, 2016.

CUTILLO, C. M.; SHARMA, K. R.; FOSCHINI, L.; KUNDU, S.; MACKINTOSH, M.; MANDL, K. D. Machine intelligence in healthcare—perspectives on trustworthiness, explainability, usability, and transparency. **Npj Digital Medicine**, v. 3, p. 37, 2020.

DEKAMIN, A.; SHEIBATOLHAMDI, A. A data mining approach for coronary artery disease prediction in Iran. **Journal of Advanced Medical Sciences and Applied Technologies**, v. *3*, n. 1, p. 29-38, 2017.

DEY, A. Machine Learning Algorithms: A Review. **International Journal of Computer Science and Information Technologies**, v. 7, n. 3, p. 1174-1179, 2016.

DING, C.; HE, X. K-means clustering via principal component analysis. In Proceedings of the twenty-first international conference on Machine learning. **International conference on Machine learning**,p. 29, 2004.

DING, S.; CONG, L.; HU, Q.; JIA, H.; SHI, Z. A multiway p-spectral clustering algorithm. **Knowledge-Based Systems**, v. 164,p.371-377, 2018.

FALUDI, A. A.; IZAR, M. C. O.; SARAIVA, J. F. K.; CHACRA, A. P. M.; BIANCO, H. T.; NETO. A. F. et al. Atualização da Diretriz Brasileira de Dislipidemias e Prevenção da Aterosclerose – 2017. **Arquivos Brasileiros de Cardiologia**, v. 109, p. 1-76, 2017.

FERNANDES, F. T.; FILHO, A. D. P. C. Data mining and machine learning perspectives for occupational safety and health. **Revista Brasileira de Saúde Ocupacional**, v. 44, p. e13, 2019.

FILHO, A. D. P. C. Uso de big data em saúde no Brasil: perspectivas para um futuro próximo. **Epidemiologia e Serviços de Saúde**, v. 24, n. 2, p.325-332, 2015.

FLORES-MATEO, G.; ROJAS-RUEDA, D.; BASORA, J.; ROS, E.; SALAS-SALVADO, J. Nut intake and adiposity: meta-analysis of clinical trials. **American Jorunal Clinical Nutrition**, v. 97, p. 1346-1355, 2013.

FRIEDEWALD, W. T.; LEVY, R. I.; FREDRICKSON, D. S. Estimation of the concentration of low-density lipoprotein cholesterol in plasma, without use of the preparative ultracentrifuge. **Clinical Chemistry**, v. 18, n. 6, p. 499–502, 1972.

GALANTE, A. P. **Desenvolvimento e validação de um método computadorizado para avaliação do consumo alimentar, preenchido por indivíduos adultos utilizando a Web.** 2007. Tese (Doutorado em Nutrição Humana Aplicada) – Faculdade de Ciências Farmacêuticas, Universidade de São Paulo, 2007.

GARCIA, L. G.; EMMENDORFER, L. R. Uma Comparação entre o algoritmo K-means e o algoritmo espectral para agrupamento de dados com curvatura acentuada. **Proceeding Series of the Brazilian Society of Applied and Computational Mathematics**, v. 5, n. 1, 2017.

GHORBANI, R.; GHOUSI, R. Predictive data mining approaches in medical diagnosis: A review of some diseases prediction. **International Journal of Data and Network Science**, v. 3, p. 47–70, 2019.

GROSS, L. S.; LI, L.; FORD, E. S.; LIU, S. Increased consumption of refined carbohydrates and the epidemic of type 2 diabetes in the United States: an ecologic assessment. **American Journal Clinical Nutrition**, v. 79, p. 774-779, 2004.

HASKEL, W. L.; LEE, I. M.; PATE, R. R.; POWELL, K. E.; BLAIR, S. N.; FRANKLIN, B. A. et al. Physical activity and publix health: updated recommendation for adults from the American college of sports medicine and the American Heart Association. **Medicine and Science in Sports and Exercise**, v. 39, n. 8, p. 1423-1434, 2007.

HORNG, S; SONTAG, D. A.; HALPERN, Y.; JERNITE, Y.; SHAPIRO, N. I.; NATHANSON, L. A. Creating an automated trigger for sepsis clinical decision support at emergency department triage using machine learning. **PLoS One**, v. 12, n. 4, p. e0174708, 2017.

JAIN, A. K. Data clustering: 50 years beyond k-means. **Pattern recognition letters**, v. 31, n. 8, p. 651–666, 2010.

JAYALATH, V. H.; DE SOUZA, R. J.; SIEVENPIPER, J. L.; HA, V.; CHIAVAROLI, L.; MIRRAHIMI. A. et al. Effect of dietary pulses on blood pressure: a systematic review and meta-analysis of controlled feeding trials. **American Journal Hypertension**, v. 27, n. 1, p. 56-64, 2013.

JOSHI, S.; NAIR, M. K. “Prediction of heart disease using classification-based data mining techniques”. **Computation al Intelligence in Data Mining**, v. 2, p. 503-511, 2015.

JUNG, Y. G.; KANG, M. S.; HEO, J. Clustering performance comparison using Kmeans and expectation maximization algorithms. **Biotechnology & Biotechnological Equipment**, v. 28, n. S1, p. S44-S48, 2014.

KARUNATHILAKE, S. P.; GANEGODA, G. U. Secondary Prevention of Cardiovascular Diseases and Application of Technology for Early Diagnosis. **Biomedicina Research International**, 2018.

KHAN, A.; BAHARUDIN, B.; LEE, L. H.; KHAN, K. A Review of Machine Learning Algorithms for Text-Documents Classification. **Journal of advances in information technology**, v. 1, n. 1, 2010.

KLUYVER, T. RAGAN-KELLEY, B.; PÉREZ, F.; GRANGER, B.; BUSSONNIER, M.; FREDERIC, J. et al. Jupyter notebooks-a publishing format for reproducible computational workflows. **In ELPUB**, p. 87–90, 2016.

KODATI, S.; VIVEKANANDAM, R.; RAVI, G. Comparative Analysis of Clustering Algorithms with Heart Disease Data sets Using Data Mining Weka Tool. **Soft Computing and Signal Processing**, p. 111-117, 2019.

KOTCHEN, T. A.; COWLEY, A. W. J. R.; FROHLICH, E. D. Salt in health and disease--a delicate balance. **New England Journal of Medicine**, v. 368, n. 13, p. 2531-2532, 2013.

KULKARNI, S. Heart Disease Classification: A Case Study using Machine Learning and Data Mining. **International Journal of Scientific Research in Computer Science, Engineering and Information Technology**, v. 4, n. 6, p. 265-271, 2018.

LI, M.; ZHEN, L.; YAO, X. How to Read Many-Objective Solution Sets in Parallel Coordinates [Educational Forum]. **IEEE Computational Intelligence Magazine**, v. 12, n. 4, 2017.

LIU, J; MAZZONE, P. J.; CATA, J. P.; KURZ, A.; BAUER, M.; MASCHA, E. J. et al. Serum free fatty acid biomarkers of lung cancer. **Chest Journal**, v. 146, n. 3, p. 670-679, 2014.

LOOK, A. R. G.; WING, R. R.; BOLIN, P.; BRANCATI, F. L.; BRAY, G. A.; CLARK, J. M. et al. Cardiovascular effects of intensive lifestyle intervention in type 2 diabetes. **New England Journal of Medicine***,* v. 369, n. 2, p. 145-154, 2013.

MA, S.; CHEN, X. A Data Mining Approach to Predict Risk of Cardiovascular. **AIP Conference Proceedings 2058**, 2019.

MALACHIAS, M. V. B.; SOUZA, W. K. S. B.; PLAVNIK, F. L.; RODRIGUES, C. I. S.; BRANDÃO, A. A.; NEVES, M. F. T. et al. VII Diretriz Brasileira de Hipertensão Arterial. **Arquivos Brasileiros de Cardiologia.** v. 107, p. 1-83, 2016.

MARTINEZ, J. C.; KING, M. P.; CAUCHI, R. Improving the Health Care System: Seven State Strategies. **National Conference of State Legislatures (NCSL)**, p. 1-28, 2016.

MARTÍNEZ-GONZÁLEZ, M. A.; FERNANDEZ-LAZARO, C.; TOLEDO, E.; DÍAZ-LÓPEZ, A.; CORELLA, D.; GODAY, A. ROMAGUERA, D. et al. Carbohydrate quality changes and concurrent changes in cardiovascular risk factors: a longitudinal analysis in the PREDIMED-Plus randomized trial. **The American Journal of Clinical Nutrition**, v. 111, n. 2, p. 291-306, 2020.

MARUCCI-WELLMAN, H. R.; CORNS, H. L.; LEHTO, M. R. Classifying injury narratives of large administrative databases for surveillance: a practical approach combining machine learning ensembles and human review. **Accident Analysis & Prevention**, v. 98, p. 359-371, 2017.

MENTE, A.; DE KONING, L.; SHANNON, H. S.; ANAND, S. S. A systematic review of the evidence supporting a causal link between dietary factors and coronary heart disease. **Archives of Internal Medicine**, v. 169, n. 7, p. 659-669, 2009.

MONTEIRO, C. A.; CANNON, G.; LEVY, R.; MOUBARAC, J-C.; JAIME, P.; MARTINS, A. P. et al. O Sistema Alimentar. **World**, v. 7, n. 1–3, 2016.

MOZAFFARIAN, D.; MICHA, R.; WALLACE, S. Effects on coronary heart disease of increasing polyunsaturated fat in place of saturated fat: a systematic review and meta-analysis of randomized controlled trials. **PLoS Medicine**, v. 7, n. 3, p.e1000252, 2010.

NESTEL, P. J.; BEILIN, L. J.; CLIFTON, P. M.; WATTS, G. F.; MORI, T. A. Practical Guidance for Food Consumption to Prevent Cardiovascular Disease. **Heart, Lung and Circulation**, v. 30, n. 2, p. 163-170, 2020.

NETO, G. B.; SILVA, E. N. D. A. Os custos da doença cardiovascular no Brasil: um breve comentário econômico. **Arquivos brasileiros de cardiologia**,v. 91, n. 4, p. 217-218, 2008.

OPAS – **ORGANIZAÇÃO PAN-AMERICANA DA SAÚDE**. XXXVI Reunión Del Comitê Asesor de Investigaciones em Salud – EncuestraMulticêntrica – Salud Beinestar y Envejecimeiento (SABE) en América Latina e el Caribe - Informe preliminar. Disponível em <URL: //WWW.opas.org/program/sabe.htm.>, mar. 2002.

PANCH, T.; SZOLOVITS, P.; ATUN, R. Artificial intelligence, machine learning and health systems**. Journal Globol Health**, v. 8, n. 2, 2018.

PARTULA, V.; DESCHASAUX, M.; DRUESNE-PECOLLO, N.; LATINO-MARTEL, P.; DESMETZ, E.; CHAZELAS, E. et al. Associations between consumption of dietary fibers and the risk of cardiovascular diseases, cancers, type 2 diabetes, and mortality in the prospective NutriNet-Santé cohort. **The American Journal of Clinical Nutrition**, v. 112, n. 1, p. 195-207, 2020.

PEDREGOSA, F.; VAROQUAUX , G.; GRAMFORT , A.; MICHEL, V.; THIRION B.; GRISEL, O. et al. Scikitlearn: Machine learning in python. **Journal of machine learning research**, p. 2825–2830, 2011.

PEREIRA, J. M. V.; CAVALCANTI, A. C. D.; SANTANA, R. F.; CASSIANO, K. M.; QUELUCI, G. C.; Guimarães T. C. F. Diagnósticos de enfermagem de pacientes hospitalizados com doenças cardiovasculares. **Anna Nery Revista de Enfermagem**, v. 15, n. 4, p. 737-745, 2011.

PHILIPI, S. T. **Tabela de composição de alimentos: suporte para decisão nutricional.**Brasília, DF, 2001.

POURIYEH, S.; VAHID, S.; SANNINO, G.; PIETRO, G; ARABNIA, H.; GUTIERREZ, J.A Comprehensive Investigation and Comparison of Machine Learning Techniques in the Domain of Heart Disease. **IEE Computer Society, Conference proceedings**, v. 1, p. 204-207, 2017.

RIDEOUT, T. C.; MARINANGELI, C.; MARTIN, H.; BROWNE, R. W, REMPEL, C. B. Consumption of low-fat dairy foods for 6 months improves insulin resistance without adversely affecting lipids or bodyweight in healthy adults: a randomized free-living cross-over study. **Nutrition Journal**, v. 12, p. 1-9,2013.

RUSSEL, S.; NORVIG, P. **Inteligência artificial**. Rio de Janeiro: Elsevier; 2013.

SCHWALM, J. D.; MCKEE, M.; HUFFMAN, M. D.; YUSUF, S. Resource Effective Strategies to Prevent and Treat Cardiovascular Disease. **Circulation**, v. 133, n. 8, p. 742–755, 2016.

SHARIFI-RAD, J.; RODRIGUES, C. F.; SHAROPOV, F.; DOCEA, A. O.; KARACA, A. C.; SHARIFI-RAD, M. et al. Diet, Lifestyle and Cardiovascular Diseases: Linking Pathophysiology to Cardioprotective Effects of Natural Bioactive Compounds. J**ournal of Environmental and Public Health**, v. 17, n. 7, p. 2326, 2020.

SHIN, J. Y.; XUN, P.; NAKAMURA, Y.; HE, K. Egg consumption in relation to risk of cardiovascular disease and diabetes: a systematic review and meta-analysis. **American Journal of Clinical Nutrition**, v. 98, n. 1, p. 146-159, 2013.

SIMENTAL-MENDÍA, L. E.; RODRÍGUEZ-MORÁN, M.; GUERRERO ROMERO, F. The product of fasting glucose and triglycerides as surrogate for identifying insulin resistance in apparently healthy subjects. **Metabolic Syndrome and Related Disorders,** v. 6, p. 299–304, 2008.

SINGH, P.; SINGH, S.; PANDI-JAI, G. S. Effective heart disease prediction system using data mining techniques. **International Journal of Nanomedicine**, v. 13, p. 121–124, 2018.

SIQUEIRA, A. S. E.; SIQUEIRA-FILHO, A. G.; LAND, M. G. P. Análise do Impacto Econômico das Doenças Cardiovasculares nos Últimos Cinco Anos no Brasil. **Arquivos Brasileiros de Cardiologia**, v. 109, n. 1, p. 39-46, 2017.

SIQUEIRA-BATISTA, R.; SILVA, E. Notas sobre os fundamentos matemáticos da Inteligência Artificial. **Revista De Ciência, Tecnologia e Inovação**, v. 4, p. 44-54, 2019.

SMITH, J. D.; HOU, T.; LUDWIG, D. S.; RIMM, E. B.; WILLETT, W.; HU, F. B. et al. Changes in intake of protein foods, carbohydrate amount and quality, and long-term weight change: results from 3 prospective cohorts. **American Journal Clinical Nutrition**, v. 101, n. 6, p. 1216-1224, 2015.

SOCIEDADE BRASILEIRA DE CARDIOLOGIA (SBC). **Arquivos Brasileiros de Cardiologia**, v. 101, p. 1-63, 2013.

STIGLIC, G.; FIJACKO, N.; KOCBEK, P.; ZITNIK, M. Interpretability of machine learning-based prediction models in healthcare. **WIREs Data Mining Knowl Discov**, e1379, 2020.

TACO **– Tabela de Composição de Alimentos**. Núcleo de Estudos e Pesquisas da UNICAMP, 2006.

TAN, P-N.; STEINBACH, M.; KUMAR, V. **Introduction to data mining. Instructor's Solution Manual**. Boston Pearson Addison-Wesley Education, 63 p., 2006.

TARAWNEH, M.; EMBARAK, O. Hybrid Approach for Heart Disease Prediction Using Data Mining Techniques. **Acta Scientific Nutritional Health**, v. 3, n. 7, p. 147-151, 2019.

UMASANKAR, P.; THIAGARASU, V. Data Mining for the Prediction of Heart Disease: A Literature Survey. **Asian Journal of Computer Science and Technology**, v. 8, n.1, p. 1-6, 2019.

USDA – United States Department of Agriculture. **Table of Nutrient Retention Factors.** Release 5, 2003.

VILLACAMPA, O. Feature selection and classification methods for decision making: A comparative analysis. PhD the sis, **College of Engineering and Computing** - Nova Southe astern University, Fort Lauderdal, 2015.

WebColorBrewer. **ProgramaColorBrewer 2.0**. Color Advice for cartography. Disponpivelem<https://colorbrewer2.org/>.

WEBER, B. FERREIRA, A.; TORREGLOSA, C.; ROSS-FERNANDES, M. B.; DA SILVA, J. T.; GALANTE, A. P. et al. The Brazilian Cardioprotective Nutritional Program to reduce events and risk factors in secondary prevention for cardiovascular disease: study protocol (The BALANCE Program Trial). **American Heart Journal**, v. 171, n. 1, p. 73–81.e2, 2016.

WITTEN, I. H.; FRANK, E.; HALL, M. **Data Mining: Practical machine learning tools and techniques.** Morgan Kaufmann, 2016.

WORLD HEALTH ORGANIZATION (WHO). **Cardiovascular diseases (CVDs) – Key Facts,** 2017.

WORLD HEALTH ORGANIZATION (WHO). **Cardiovascular diseases (CVDs)**, 2020. Disponível em<http://origin.who.int/cardiovascular_diseases/en/> 2020. Acessado em Setembro de 2020.

WORLD HEALTH ORGANIZATION (WHO). **Cardiovascular diseases (CVDs)**, 2021. Disponível em <https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds). Acessado em Julho de 2021.

WORLD HEALTH ORGANIZATION (WHO). **Waist circumference and waist–hip ratio: report of a WHO expert consultation**. 2008.

WORLD HEART FEDERATION, 2019. https://www.world-heart-federation.org/world-heart-day/world-heart-day-2019/. Acessado em 9 de Setembro de 2021.

YANG, Y.; PEDERSEN, J. O. A comparative study on feature selection in text categorization. **International Conference on Machine Learning**, v. 97, p. 412-420, 1997.

YILDIZ, B.; BILBAO, J. I.; SPROUL, A. B. A review and analysis of regression and machine learning models on commercial building electricity load forecasting. **Renewable and Sustainable Energy Reviews,** v. 73, p. 1104-1122, 2017.

ZHENG, J.; HUANG, T.; YU, Y.; HU, X.; YANG, B.; LI, D. Fish consumption and CHD mortality: an updated meta-analysis of seventeen cohort studies. **Public Health Nutrition**, v. 15, n. 4, p. 725-737, 2012.

ZHENG, Q.; DELINGETTE, H.; FUNG, K.; PETERSEN, S. E.; AYACHE, N. Unsupervised shape and motion analysis of 3822 cardiac 4D MRIs of UK Biobank. **Preprint submitted to arXiv**, 2019.

8 ANEXO

8.1 ANEXO I - Parecer consubstanciado do Comitê de Ética e Pesquisa com Seres Humanos do Hcor

HOSPITAL DO CORAÇÃO/ ASSOCIAÇÃO DO SANATÓRIO SÍRIO - ASS

PARECER CONSUBSTANCIADO DO CEP

DADOS DO PROJETO DE PESQUISA

Título da Pesquisa: Efeito do Programa Alimentar Brasileiro Cardioprotetor na redução de eventos e fatores de risco na prevenção secundária para doença cardiovascular: um ensaio clínico randomizado

Pesquisador: Bernardete Weber

Área Temática:

Versão: 26

CAAE: 03218512.0.1001.0060

Instituição Proponente: Hospital do Coração/ Associação do Sanatório Sírio

Patrocinador Principal: Hospital do Coração/ Associação do Sanatório Sírio

DADOS DO PARECER

Número do Parecer: 1.171.748

Data da Relatoria: 14/07/2015

Apresentação do Projeto:

BRAZILIAN CARDIOPROTECTIVE DIET TRIAL - Efeito do Programa Alimentar Brasileiro Cardioprotetor na redução de eventos e fatores de risco na prevenção secundária para doença cardiovascular. Serão recrutados cerca de 2000 pacientes, em 40 centros distribuídos no Território Nacional, por recrutamento competitivo.

Objetivo da Pesquisa:

Primário: avaliar a efetividade do Programa Alimentar Brasileiro Cardioprotetor (PABC) na redução de:parada cardíaca, infarto agudo do miocárdio, AVC, revascularização do miocárdio, amputação por doença arterial periférica, angina ou óbito.

Secundário: avaliar a efetividade do plano alimentar na redução de fatores de risco (colesterol total, LDL, glicemia, pressão arterial, IMC, circunferência da cintura).

Avaliação dos Riscos e Benefícios:

Risco mínimo. A pesquisa não envolve riscos potenciais aos participantes, visto que as recomendações de dietoterapia para prevenção e tratamento de doenças cardiovasculares são condutas consagradas pela comunidade científica. Por ser inédita e contar com regionalização da alimentação, a pesquisa pode contribuir com a definição de uma dieta brasileira cardioprotetora de fácil execução em nível nacional.

Endereço: Rua Abrão Dib, 50 - Térreo
Bairro: Paraíso **CEP:** 04.004-030
UF: SP **Município:** SAO PAULO
Telefone: (11)3886-4688 **Fax:** (11)3886-4689 **E-mail:** etica.pesquisa@hcor.com.br

Comentários e Considerações sobre a Pesquisa:

Referente ao estudo supramencionado submetemos para apreciação o protocolo versão 5 com a nova versão do TCLE (TCLEdieta_Cardio_7.2), ambos aprovados pelo CEP-HCor em 11/06/2015 e 19/01/2015, respectivamente. As versões desses documentos dizem respeito ao prolongamento do tempo de seguimento sem alteração nos objetivos do estudo ou dados coletados.

O novo TCLE será aplicado na consulta de 18 meses, nos pacientes que já participam do estudo, com o intuito de convidá-los a participar por mais 36 meses.

Alguns participantes consentem em participar do prolongamento do estudo, porém não estão dispostos em ir (se deslocar) até o centro colaborador para a coleta de dados. Por outro lado, não se opõem a receber ligações telefônicas periódicas a fim de coletar dados referentes à sua saúde. Isto é, eles consentem em nos passar as informações, porém não gostariam de se deslocar para a consulta, o que dificulta a obtenção da assinatura do TCLE.Desta forma, uma vez que não há inclusão da coleta de novos dados no contato telefônico, foi solicitado avaliação ética sobre a possibilidade de não-reconsentimento desses participantes que optam pelo acompanhamento exclusivo via telefone (à distância).

Considerações sobre os Termos de apresentação obrigatória:

As alterações referidas na emenda não apresentam restrição ética.

Recomendações:

Sem recomendações.

Conclusões ou Pendências e Lista de Inadequações:

Sem pendências.

Situação do Parecer:

Aprovado

Necessita Apreciação da CONEP:

Não

8.2 ANEXO II - Parecer consubstanciado do Comitê de Ética em Pesquisa com Seres Humanos da Universidade Federal de Viçosa

UNIVERSIDADE FEDERAL DE VIÇOSA - UFV

PARECER CONSUBSTANCIADO DO CEP

DADOS DO PROJETO DE PESQUISA

Título da Pesquisa: Efeito do Programa Alimentar Brasileiro Cardioprotetor na redução de eventos e fatores de risco na prevenção secundária para doença cardiovascular: um ensaio clínico randomizado

Pesquisador: Josefina Bressan

Área Temática:

Versão: 3

CAAE: 03218512.0.2002.5153

Instituição Proponente: Universidade Federal de Viçosa - UFV

Patrocinador Principal: Hospital do Coração/ Associação do Sanatório Sírio

DADOS DO PARECER

Número do Parecer: 1.020.056

Data da Relatoria: 10/04/2015

Apresentação do Projeto:

O projeto já foi aprovado por este Comitê com parecer número 882.612 de 17/11/14. Trata-se de uma emenda ao referido projeto estendendo o tempo de acompanhamento dos pacientes de 12 para 48 meses.

Objetivo da Pesquisa:

Primário: avaliar a efetividade do Programa Alimentar Brasileiro Cardioprotetor (PABC) na redução de: parada cardíaca, infarto agudo do miocárdio, AVC, revascularização do miocárdio, amputação por doença arterial periférica, angina ou óbito. Secundário: avaliar a efetividade do plano alimentar na redução de fatores de risco (colesterol total, LDL, glicemia, pressão arterial, IMC, circunferência da cintura).

Avaliação dos Riscos e Benefícios:

Riscos e benefícios descritos de acordo com a Resolução 466/12.

Comentários e Considerações sobre a Pesquisa:

A emenda implica em adequações ao projeto que foram devidamente apresentadas e justificadas.

Endereço: Universidade Federal de Viçosa, Edifício Arthur Bernardes, piso inferior
Bairro: Campus Universitário **CEP:** 36.570-000
UF: MG **Município:** VICOSA
Telefone: (31)3899-2492 **Fax:** (31)3899-2492 **E-mail:** cep@ufv.br

Considerações sobre os Termos de apresentação obrigatória:

Termos apresentados de acordo com a Resolução 466/12.

Recomendações:

Conclusões ou Pendências e Lista de Inadequações:

Aprovado pedido de emenda.

Situação do Parecer:

Aprovado

Necessita Apreciação da CONEP:

Não

Considerações Finais a critério do CEP:

VICOSA, 13 de Abril de 2015

Assinado por:
Patricia Aurélia Del Nero
(Coordenador)

8.3 ANEXO III - Termo de Consentimento Livre e Esclarecido

ESTUDO DIETA CARDIOPROTETORA
TERMO DE CONSENTIMENTO LIVRE E ESCLARECIDO

INTRODUÇÃO

Dados recentes da Organização Mundial de Saúde (OMS) demonstram que as doenças do coração representam a principal causa de morte no Brasil e no Mundo. Estudos demonstram que a alimentação adequada pode diminuir o risco para a maioria dos casos de doença do coração como, por exemplo, infarto e derrame.

O Brasil é um país tropical, muito rico em alimentos considerados saudáveis, como frutas, verduras, legumes e grãos. Elaborar um Programa Alimentar que valorize os alimentos brasileiros e respeite as diferenças culturais entre as regiões do Brasil pode ser fundamental para diminuir estas doenças.

Estudos de avaliação do efeito de uma dieta/orientação nutricional na prevenção de doenças cardiovasculares apresentam tempo de seguimento de pelo menos três anos. Isto é, para avaliar o efeito da dieta parece ser necessário acompanhar os pacientes sob orientação, por um período de pelo menos três anos.

OBJETIVOS

É por isso que estamos propondo aumentar o tempo de seguimento desta pesquisa que tinha, a princípio, um tempo de seguimento de 12 meses, para um seguimento de até 48 meses. Após este período, avaliaremos o efeito do Programa Alimentar Brasileiro Cardioprotetor, na redução de doenças cardíacas e também do colesterol, as gorduras no sangue, glicemia, o açúcar do sangue, a pressão alta e a obesidade.

Por verificarmos a necessidade de ampliar o tempo de acompanhamento, é que o(a) senhor(a) esta sendo convidado(a) a continuar no estudo por mais um tempo, até novembro de 2017.

PROCEDIMENTOS DO ESTUDO / COMO É O ESTUDO?

COMO É O ESTUDO?

Se o(a) senhor(a) estiver no grupo A, o(a) senhor(a) continuará recebendo orientação do Programa Alimentar Brasileiro Cardioprotetor. A forma de acompanhamento será a seguinte:

- Até 24 meses (dois anos), consultas a cada seis meses.
- Após 24 meses (dois anos), o(a) senhor(a) passará a ter além de consultas individuais encontros em grupo com os outros participantes da pesquisa. Assim, o contato passará a ser a cada 4 meses.

Nas consultas será solicitado ao senhor que realize um exame de sangue, será medido seu peso, a circunferência da sua cintura, e também serão feitas perguntas sobre o que o(a) senhor(a) comeu no dia anterior à consulta. O objetivo dos encontros em grupo será de incentivar a troca de experiências entre os participantes e estimular a adesão à dieta. Além dessas visitas presenciais, o(a) senhor(a) também receberá ligações de pesquisadores do centro coordenador deste estudo, que fica em São Paulo/SP, e que tem objetivo de verificar como está sua saúde (por exemplo, se o(a) senhor(a) mudou a medicação) e esclarecer dúvidas da orientação nutricional. Essas ligações são realizadas nos meses em que o(a) senhor(a) não tem contato presencial com o pesquisador.

Se o(a) senhor(a) estiver no grupo B, o(a) senhor(a) continuará recebendo orientação da sua alimentação específica para suas necessidades. A forma de acompanhamento será a seguinte:

- Até 48 meses, consultas a cada seis meses.

Nas consultas será solicitado ao senhor que realize um exame de sangue, será medido seu peso, a circunferência da sua cintura, e também serão feitas perguntas sobre o que o(a) senhor(a) comeu no dia anterior à consulta. Além dessas visitas presenciais, o(a) senhor(a) também receberá ligações de pesquisadores do centro coordenador deste estudo, que fica em São Paulo/SP, e que tem objetivo de verificar como está sua saúde (por exemplo, se o(a) senhor(a) mudou a medicação) e esclarecer dúvidas da orientação nutricional. Essas ligações são realizadas nos meses em que o(a) senhor(a) não tem contato presencial com o pesquisador.

DESCONFORTOS E RISCOS

Talvez o(a) senhor(a) sinta um pequeno desconforto no momento da coleta de sangue, devido a picada de agulha da seringa. Mas este desconforto é momentâneo. Também não sofrerá nenhum risco em participar do projeto, pois as orientações nutricionais de ambos os grupos (A e B) já são validada por diversas sociedades médicas brasileiras e internacionais, o que estamos prevendo é apenas a forma de orientar, de abordar o tema com o paciente.

BENEFÍCIOS

Como benefícios, o senhor poderá ter seu colesterol, gordura e açúcar no sangue, peso e pressão do sangue reduzidos, e terá acesso aos resultados de todos os exames.

Sua participação é totalmente voluntária e o (a) senhor (a) pode desistir e retirar seu consentimento em qualquer momento durante o decorrer da pesquisa, sem que isso prejudique sua assistência pela equipe de saúde.

EFEITOS ADVERSOS

Não é previsto efeitos adversos ao estudo, visto que a composição da dieta já é validada. Estamos apenas comparando duas formas de orientação da dieta, a forma didática de se prescrever algo já consolidado. De qualquer forma, está garantido seu direito de indenização diante de eventuais danos decorrente da intervenção.

OUTRAS INFORMAÇÕES

O (a) senhor (a) não terá nenhum custo por participar da pesquisa. Os gastos com transporte, exames laboratoriais e lanche (nos dias dos exames de sangue), serão ressarcidos pela pesquisa.

Seus dados são secretos e sigilosos de acordo com as normas brasileiras durante todas as fases da pesquisa. Os resultados desta pesquisa poderão ser publicados em revistas científicas, mas a sua identidade será preservada.

A qualquer momento o(a) senhor(a) poderá esclarecer dúvidas através dos seguintes contatos:

- Dra Bernardete Weber – Investigadora Principal: fone (xx-11) 3053-6611 (ramal 1124).
- Dr Otávio Berwanger – Investigador Responsável: fone (xx-11) 3053-6611 (ramal 8201).
- Comitê de Ética em Pesquisa do Hospital do Coração: fone (xx-11) 3886-4688. Para esclarecimentos sobre aspectos éticos do estudo. Rua Abrão Dib, 50 – Térreo – São Paulo, SP
- Contato por e-mail: projetodicabr@gmail.com

COMO PARTICIPAR?

A participação neste estudo é inteiramente voluntária. Para isso o(a) senhor(a) deve assinar esse Termo de Consentimento Livre e Esclarecido. O senhor receberá uma cópia deste termo, com todas as informações e contatos dos pesquisadores. O(a) senhor(a) poderá deixar de participar do estudo em qualquer fase da pesquisa, sem penalização alguma, poderá retirar seu consentimento em qualquer momento.

Declaro que li o termo de consentimento livre e esclarecido para esse estudo e aceito participar voluntariamente desse estudo. Ainda, declaro que recebi todos os esclarecimentos necessários para compreender o estudo e tive tempo suficiente para decidir minha participação no estudo.

Nome do Paciente: _____

(ou representante legal)
Assinatura do Paciente:_____**Data:**_____
(ou representante legal)

Investigador: _____

Assinatura:_____**Data:**_____

8.4 ANEXO IV - Ficha clínica da visita inicial

Dieta cardioprotetora

DIETA CARDIOPROTETORA BRASILEIRA

BRAZILIAN CARDIOPROTECTIVE DIET TRIAL

EFEITO DO PROGRAMA ALIMENTAR CARDIOPROTETOR NA REDUÇÃO DE EVENTOS E FATORES DE RISCO NA PREVENÇÃO SECUNDÁRIA PARA DOENÇA CARDIOVASCULAR: UM ENSAIO CLÍNICO RANDOMIZADO

FICHA CLÍNICA
GRUPO INTERVENÇÃO

Iniciais do paciente ▢▢▢

N° de Identificação ▢▢ ▢▢▢▢

n° do centro n° do paciente

Hospital do Coração HCor São Paulo

Dieta cardioprotetora

DADOS BASAIS

Identificação ◯◯ n° do centro ◯◯◯◯ n° do paciente

Iniciais do paciente ◯◯◯

DATA ◯◯/◯◯/◯◯

1. Antropometria

Peso 1(Kg) ◯◯◯,◯ Altura (m.) ◯,◯◯◯

Peso 2(Kg) ◯◯◯,◯ IMC (Kg/m²) ◯◯,◯

Peso médio(kg) ◯◯◯,◯

Circunferência da cintura 1 (cm) ◯◯◯,◯

Circunferência da cintura 2 (cm) ◯◯◯,◯

Circunferência da cintura média (cm) ◯◯◯,◯

1. Aferição da pressão arterial

Pressão arterial sistólica (mmHg) ◯◯◯

Pressão arterial diastólica (mhg) ◯◯◯

3. Exames bioquímicos

Colesterol total (mg/dl) ◯◯◯

LDL colesterol (mg/dl) ◯◯◯

HDL colesterol (mg/dl) ◯◯◯

Triglicérides (mg/dl) ◯◯◯◯

Glicemia de jejum (mg/dl) ◯◯◯

4. Medicação

Anticoagulantes

AAS	◯Sim ◯Não
Clopidogrel	◯Sim ◯Não
Prasugrel	◯Sim ◯Não
Ticagrelor	◯Sim ◯Não
Warfarina	◯Sim ◯Não
Outro anticoagulante	◯Sim ◯Não

Especificar _____

Anti hipertensivos

Beta-bloqueadores	◯Sim ◯Não
Diurético tiazídico	◯Sim ◯Não
Inibidor de ECA	◯Sim ◯Não
Bloq. receptor Angiotensina II	◯Sim ◯Não
Bloqueador de Renina	◯Sim ◯Não
Antagonista de canais de cálcio	◯Sim ◯Não

Antilipemiantes e estatina ◯Sim ◯Não

Hipoglicemiantes

Metformina	◯Sim ◯Não
Insulina	◯Sim ◯Não
Outro hipoglicemiante oral	◯Sim ◯Não

Especificar: _____

Observações:

CRF preenchido por: _____ Data: _____

Hospital do Coração HCor São Paulo

Dieta cardioprotetora

DADOS BASAIS

Identificação ☐☐ ☐☐☐☐ Iniciais do paciente ☐☐☐

nº do centro nº do paciente

5. Características clínicas

Hipertensão arterial ☐Sim ☐Não

Diabetes Mellitus ☐Sim ☐Não

Dislipidemia ☐Sim ☐Não

História familiar de DAC ☐Sim ☐Não

Fumante ☐ Ex-fumante ☐ Nunca fumou ☐

Com que idade começou a fumar ☐☐

Números de cigarros/dia ☐☐

Com que idade parou de fumar ☐☐

Frequência de exposição a outras pessoas fumando Nunca ☐ 1-2x/sem ☐ 3-6x/sem ☐ Diário ☐

6. Fator Atividade Física

☐ Sedentário
☐ Atividade leve
☐ Atividade média
☐ Atividade alta

CRF preenchido por: _____ Data: _____

Hospital do Coração HCor São Paulo

RECORDATÓRIO DE 24H ALIMENTAR

Dieta cardioprotetora

Recordatório _____ Data _____

Pesquisador _____

QUE DIA DA SEMANA FOI ONTEM? (PESQUISADOR: Você é quem deve responder esta questão, não solicite a resposta ao entrevistado)

( ) SEGUNDA ( ) TERÇA ( ) QUARTA ( ) QUINTA ( ) SEXTA ( ) SÁBADO ( ) DOMINGO

HORÁRIO	REFEIÇÃO	ALIMENTO preparação/bebida	QUANTIDADE Medidas caseiras	MARCA COMERCIAL	OBSERVAÇÕES

8.5 ANEXO V - Ficha clínica da visita de 15 dias

Hospital do Coração HCor São Paulo

Dieta cardioprotetora

VISITA CLÍNICA DE 15 DIAS

Identificação ◯◯ ◯◯◯◯ Iniciais do paciente ◯◯◯

nº do centro nº do paciente

DATA ◯◯/◯◯/◯◯

Compareceu à visita? ◯Sim ◯Não

Se não, qual a razão?

◯ Óbito Data do óbito ___/___/___
◯ Óbito Cardiovascular
◯ Desistência
◯ Esquecimento
◯ Internação

1. Medicação

Anticoagulantes

AAS	◯Sim ◯Não
Clopidogrel	◯Sim ◯Não
Prasugrel	◯Sim ◯Não
Ticagrelor	◯Sim ◯Não
Warfarina	◯Sim ◯Não
Outro anticoagulante	◯Sim ◯Não

Especificar _____

Observações:

Anti hipertensivos

Beta-bloqueadores	◯Sim ◯Não
Diurético tiazídico	◯Sim ◯Não
Inibidor de ECA	◯Sim ◯Não
Bloq. receptor Angiotensina II	◯Sim ◯Não
Bloqueador de Renina	◯Sim ◯Não
Antagonista de canais de cálcio	◯Sim ◯Não
Antilipemiantes e estatina	◯Sim ◯Não

Hipoglicemiantes

Metformina	◯Sim ◯Não
Insulina	◯Sim ◯Não
Outro hipoglicemiante oral	◯Sim ◯Não

Especificar: _____

2. Valor Energético Recomendado (Kcal)

◯◯◯◯

Observações:

CRF preenchido por: _____ Data: _____

Hospital do Coração HCor São Paulo

RECORDATÓRIO DE 24H ALIMENTAR

Dieta cardioprotetora

Recordatório _____ Data _____

Pesquisador _____

QUE DIA DA SEMANA FOI ONTEM? (PESQUISADOR: Você é quem deve responder esta questão, não solicite a resposta ao entrevistado)

( ) SEGUNDA ( ) TERÇA ( ) QUARTA ( ) QUINTA ( ) SEXTA ( ) SÁBADO ( ) DOMINGO

HORÁRIO	REFEIÇÃO	ALIMENTO preparação/bebida	QUANTIDADE Medidas caseiras	MARCA COMERCIAL	OBSERVAÇÕES