

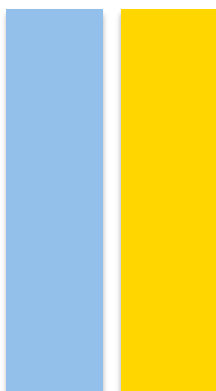
2º CICLO DE ESTUDOS
MESTRADO EM INFORMÁTICA MÉDICA

Data Mining approaches for Bio-Demographic and Morbidity analysis

Pedro Manuel de Ascensão Rocha Vieira Nogueira

M

2024



2º CICLO DE ESTUDOS
MESTRADO EM INFORMÁTICA MÉDICA

Data Mining approaches for Bio-Demographic and Morbidity analysis

Pedro Manuel de Ascensão Rocha Vieira Nogueira

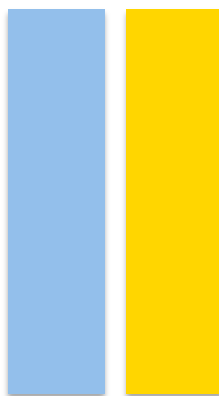
Orientadores:

Professor Doutor Hélder Pinto Oliveira

Professor Doutor Pedro Pereira Rodrigues

M

2024



“We can only see a short distance ahead, but we can see plenty there that needs to be done”

Alan Turing

Sumário

Este trabalho, abordou de uma forma estruturada o *data mining* de uma base de dados de saúde, demonstrando que, com uma abordagem simples é possível extrair informações clínicas relevantes sobre a população em estudo. O principal objetivo era caracterizar uma população, sob ponto de vista biodemográfico e perfil de morbilidades (TOP 10 das patologias mais prevalentes), por concelho de residência. A população alvo deste estudo foram os utentes adultos (≥ 18 anos) que recorreram ao Grupo Lusíadas Saúde. Realizou-se um estudo observacional transversal com o trabalho numa base de dados de registos eletrónicos médicos. A recolha de dados proporcionou a análise e caracterização da população que recorreu pelo menos uma vez à consulta programada ou não programada, entre 1 de agosto 2022, até 1 agosto 2023.

Focado na extração de informação clínica relevante no grupo estudado foram demonstradas algumas oportunidades de aplicação prática da informação das patologias e comorbilidades relevantes. Também foi delimitada as áreas de intervenção através dos 4 dígitos do código postal. Através da clusterização por Kmeans demonstra-se que é possível obter informação relevante mostrando como oportunidade para desenvolver um plano de ação para as equipas multidisciplinares sem comprometer requisitos éticos ou de proteção de dados.

Espera-se que este estudo possa ser encarado como um ponto de partida para a otimização de planos de saúde às populações alvo com fundamentação em informações clínicas retiradas de base de dados de qualidade.

Abstract

This work addressed the data mining of a health database in a structured way, demonstrating that with a simple approach, it is possible to extract relevant clinical information about the studied population. The main objective was to characterize a population from a biodemographic perspective and in terms of morbidity profiles (TOP 10 most prevalent pathologies), categorized by municipality of residence. The target population of this study consisted of adult patients (≥ 18 years) who utilized the Lusíadas Saúde Group services.

A cross-sectional observational study was conducted using a database of electronic medical records. Data collection enabled the analysis and characterization of the population who attended at least one scheduled or unscheduled consultation between August 1, 2022, and August 1, 2023.

Focusing on extracting relevant clinical information from the studied group, several opportunities for practical applications of the data on significant pathologies and comorbidities were demonstrated. Intervention areas were also delineated using the four-digit postal code. Through K-means clustering, it was shown that obtaining relevant information is feasible, highlighting opportunities to develop action plans for multidisciplinary teams without compromising ethical or data protection requirements.

It is hoped that this study can serve as a starting point for optimizing healthcare plans for target populations, based on clinically substantiated information extracted from high-quality databases.

Índice

Sumário	iv
Abstract	vi
Índice.....	viii
Índice de Figuras	ix
Índice de Tabelas	x
Acrónimos	xi
1. Introdução	12
1.1 Disability Adjusted Life Years	13
1.2 Intervenções Clínicas	14
1.3 Objectivo	16
2. Fundamentação teórica	17
2.1 <i>Data Mining</i> e Registos Eletrónicos Saúde	18
2.2 Estudos nacionais com foco em DALYs	28
3. Metodologia.....	32
4. Resultados e discussão	35
4.1 Caracterização da população em estudo	37
4.2 Caracterização da população em estudo após filtragem.....	41
4.3 Oportunidades na intervenção clínica.....	44
4.4 Resultados com abordagem pela inteligência computacional.....	47
4.5 Discussão	53
5. Conclusão.....	55
6. Referências	56
7. Anexos	63

Índice de Figuras

Figura 1	Gráfico do Top 10 ICD9 na base de dados Norte	38
Figura 2	Gráfico do Top 10 ICD9 na base de dados Centro	38
Figura 3	Gráfico do Top 10 ICD9 na base de dados Sul	38
Figura 4	Gráfico das DALYs (%) nos doentes Lusíadas	41
Figura 5	Gráfico com áreas das comorbilidades mais comuns da doença isquémica no grupo estudado	45
Figura 6	Exemplo das primeiras linhas da BD Norte com a variável dicotómica	46
Figura 7	Matriz de correlação das patologias na BD Norte	47
Figura 8	Gráfico de análise silhueta	48
Figura 9	Gráfico PCA – visualização de clusters	49

Índice de Tabelas

Tabela 1	Ordenação em 2019 da taxa de DALYs em Portugal	12
Tabela 2	Descrição das bases de dados – considerando adultos (≥ 18 anos)	36
Tabela 3	Correspondência DALYs – Código ICD9	39
Tabela 4	Descrição das bases de dados filtrado por DALYs	41
Tabela 5	Centros clusters	47

Acrónimos

APIs	<i>Application Programming Interfaces</i>
ACM	<i>Association for computing machinery</i>
CRAN	<i>Comprehensive R Archive Network</i>
DALYs	<i>Disability-Adjusted Life Years</i>
DPOC	Doença Pulmonar Obstrutiva Crónica
DNTs	Doenças Não Transmissíveis
GBD	<i>Global Burden of Disease</i>
IMC	Índice de Massa Corporal
IEEE	<i>Institute of Electrical and Electronics Engineers</i>
IC	Inteligência Computacional
ICD	<i>Internacional Classification of Disease</i>
IoT	<i>Internet of Things</i>
OMS	Organização Mundial de Saúde
PCA	<i>Principal componente analysis</i>
RES	Registos Eletrónicos de Saúde
RGPD	Regulamento Geral sobre a Proteção de Dados

1. Introdução

Os conjuntos de dados que podem ser extraídos em saúde estão a expandir-se, permitindo a criação de imagens cada vez mais detalhadas dos doentes e do seu estilo de vida. É importante o desenvolvimento de ferramentas que permitam extrair conhecimento para desenvolver-se intervenções direcionadas para a melhoria da qualidade de vida das populações.^[1]

Após a leitura de um artigo da Harvard Business Review e enquanto profissional de saúde com vinte anos de experiência clínica alertou-me, a seguinte constatação, que 50% dos recursos são para 5% dos doentes. Dessa pequena percentagem identificam-se 3 grupos de doentes:

1. doentes sujeitos a um problema de saúde catastrófico, por exemplo um acidente, neoplasia ou situação muito grave aguda.
2. condições médicas críticas como insuficiência cardíaca descompensada ou doença renal crônica entre outras, que requerem tratamento contínuo escalável. Dificilmente se altera a situação potenciando grandes despesas em cuidados de saúde no futuro.
3. uma ou mais doenças crónicas que podem melhorar e serem estabilizadas - geralmente condições crónicas como diabetes, insuficiência cardíaca estável, asma e doenças mentais. Estas

doenças levam os doentes frequentemente às urgências, a consultas médicas e hospitalizações periódicas. A capacitação dos doentes para a gestão da sua patologia pode ser muito eficaz, através de ações simples e de baixo custo, levar ao aumento da qualidade de vida e à redução dos recursos de saúde necessários com a diminuição do agravamento destas patologias, retirando-os do grupo dos grandes consumidores de recursos de cuidados de saúde^[2]

1.1 Disability Adjusted Life Years

Para este terço de doentes em que se pretende intervir pode-se considerar uma abordagem sobre a taxa de Anos de Vida Ajustados por Incapacidade ou DALYs (Disability Adjusted Life Years) que é a taxa da mortalidade (anos de vida perdidos devido à morte prematura) com a morbilidade (anos vividos com uma doença ou incapacidade).

Tabela 1 Ordenação em 2019 da taxa de DALYs em Portugal. ^[3]

1º	Doença cerebrovascular
2º	Doença isquémica do coração
3º	Dores lombares
4º	Diabetes <i>mellitus</i>
5º	DPOC
6º	Depressão
7º	Tumor maligno do pulmão
8º	Infeções respiratórias inferiores
9º	Tumor maligno do cólon e reto
10º	Doença Alzheimer

Ao quantificar o peso relativo de diferentes condições de saúde, as DALYs permitem uma comparação objetiva e sistemática do impacto de diferentes doenças. Analisando o Top nacional, as DNTs (doenças não transmissíveis) têm um impacto significativo, observam-se doenças cardiovasculares, diabetes, obesidade e cancro, que representam uma parcela significativa da mortalidade também na Europa. Fatores como tabagismo, alimentação inadequada, sedentarismo e consumo de álcool contribuem para essas doenças. As ações clínicas focadas em medidas preventivas, detecção precoce e tratamentos eficazes podem ter um impacto significativo na redução das DNTs, melhorando a saúde das populações e a eficiência dos sistemas de saúde, alinhando-se com o objetivo de proporcionar mais valor e qualidade nos cuidados de saúde oferecidos.^[4]

Além disso, a análise das DALYs pode fornecer uma base sólida para a alocação eficiente de recursos de saúde. Identificar as doenças e lesões que causam a maior carga de doença permitindo direcionar recursos para áreas prioritárias, como prevenção, diagnóstico precoce, tratamento adequado e reabilitação. Isso contribui para melhorar os resultados em saúde e maximizar o impacto positivo das intervenções de saúde.

1.2 Intervenções Clínicas

O processo transicional que o doente e seus familiares atravessam não se esgota no episódio na unidade de saúde, muito pelo contrário, apenas é o ponto de partida. Também para uma sociedade equilibrada e responsável estes doentes são um desafio, na procura de equidade, acessibilidade aos cuidados e a melhor qualidade de vida possível.

A literacia em saúde desempenha um papel fundamental na capacitação dos indivíduos a fim de tomar decisões informadas sobre si. Ao promover uma compreensão abrangente, incluindo conceitos médicos complexos, acesso a serviços de saúde e a habilidade de avaliar criticamente informações relacionadas à saúde, a literacia em saúde pode capacitar os indivíduos a adotar comportamentos saudáveis e procurar cuidados de saúde adequados.^[5]

Nesse sentido, as intervenções breves, conforme propostas pela Organização Mundial de Saúde (OMS), emergem como uma estratégia eficaz para a promoção da saúde e prevenção de doenças em diversos contextos de atendimento de saúde. Essas intervenções são baseadas em evidências científicas e têm como objetivo fornecer informações claras, concisas e personalizadas aos indivíduos, de forma a melhorar sua compreensão sobre comportamentos saudáveis e riscos à saúde. Ao adotar uma abordagem centrada no paciente, estas intervenções têm o potencial de melhorar a literacia em saúde. Esta abordagem para capacitar os indivíduos, promove a sua saúde sendo essencial para o fortalecimento do sistema de saúde, com o foco numa população mais informada e capaz de tomar medidas proativas em relação à sua própria saúde.^[6,7]

Também as consultas de grupo devem ser levadas em conta dado que proporciona um ambiente seguro e de apoio, os participantes partilham experiências, recebem apoio emocional e desenvolvem competências. Além disso, estas intervenções promovem a coesão do grupo e encorajam a aprendizagem social e a troca de perspetivas. A interação com pares que enfrentam desafios semelhantes cria um sentimento de pertença e reduz o isolamento e o estigma. De realçar a necessidade de avaliar cuidadosamente quais condições de saúde e quais os doentes que poderiam beneficiar

levando em consideração as necessidades e preferências individuais.^[8] A disponibilização de recursos locais e da comunidade onde vive o indivíduo, seriam custos menores, para englobar num plano de intervenção clínica e com a potencialidade de serem grandes poupadores de recursos de saúde a longo prazo.

1.3 Objectivo

A análise bio-demográfica e o estudo das morbilidades são elementos essenciais para compreender a saúde da população e desenvolver estratégias eficazes e políticas de saúde. No campo da epidemiologia, a caracterização da população sob diferentes aspetos, como idade, sexo, condições socioeconômicas e perfil de doenças, desempenha um papel crucial. Assim o objetivo principal é caracterizar a população adulta utilizadora, pelo menos 1 vez ao longo de 12 meses, das unidades Lusíadas Saúde em Portugal Continental sob o ponto de vista biodemográfico e perfil de morbilidades (Top 10).

Como objetivos secundários:

- identificar todos os indivíduos, dentro da população, que tenham pelo menos um diagnóstico coincidente com a lista TOP 10 de Portugal (de 2019) de Disability-adjusted life years (DALYs);
- determinar a prevalência de multimorbilidade, nos indivíduos que tenham pelo menos um diagnóstico coincidente com a lista nacional de TOP 10 de DALYs;
- estudar a distribuição dos indivíduos com DALYs pelos diferentes códigos postais (4 dígitos iniciais), em cada uma das regiões (norte, centro e sul).

2. Fundamentação teórica

A fundamentação teórica situa a investigação no contexto das pesquisas mais recentes, identificando tendências, lacunas e questões não resolvidas. Permite a compreensão dos termos e as ideias-chave utilizadas ao longo do trabalho. Justifica a relevância da pesquisa ao demonstrar como ela se relaciona e assegura que os métodos utilizados são apropriados e robustos, alinhando-se com os objetivos da pesquisa

Este capítulo, encontra-se dividido em dois subcapítulos cada um deles apresentando a sua *query* assim como a formulação de termos de pesquisa específicos baseados nos objetivos do estudo; a seleção das fontes com a escolha de bases de dados relevantes na área de informática médica; revisão dos resultados com a leitura e análise crítica dos artigos encontrados, destacando tendências, desafios e contextualizando este trabalho com as pesquisas existentes. Para a realização deste capítulo foram realizadas duas *queries*. Na primeira pesquisa pretende-se conhecer o estado da arte focado no que é o *data mining* de uma base de dados retirada de registos eletrónicos de saúde. Numa segunda *query* procurou-se pesquisar sobre o estado da arte para estudos na população portuguesa com uma abordagem que poderia ter semelhanças com este estudo e que relacionavam, características biodemográficas, *data mining* e DALYS,

2.1 *Data Mining* e Registos Eletrónicos Saúde

Com esta primeira revisão pretende-se conhecer os desafios levantados evitando os erros já identificados por outros investigadores. Seguir as recomendações dos artigos estudados de forma a realizar um trabalho bem alicerçado no conhecimento da área. Assim e de encontro com a pergunta levantada foi colocada a seguinte query: ("Epidemiologic Methods" OR "Data Collection" OR "Vital Statistics") AND ("electronic health records" OR "health care") AND ("Data Mining" OR "CRISP-DM" OR "Cross Industry Standard Process for Data Mining").

A pesquisa foi realizada a 19 de outubro de 2023 na PUBMED, na biblioteca digital do *Institute of Electrical and Electronics Engineers* (IEEE) e na da *Association for computing machinery* (ACM) com os seguintes critérios:

- de inclusão - artigos em língua portuguesa ou inglesa, publicados entre 2022-2023;
- de exclusão - documentos de conferencias, artigos de opinião, cartas ao editor.

Daí resultaram 18 artigos da biblioteca digital da ACM, 12 artigos na do IEEE e 4 artigos na PUBMED.

Foi recolhido um background importante para a estruturação da abordagem a realizar à base de dados deste trabalho. Com as especificações do âmbito da saúde e com o intuito de criação de informação de valor com a utilização de ferramentas de *data mining*.

A tecnologia da informação em saúde resulta atualmente numa extensa recolha de grandes volumes de dados. Os cuidados de saúde prestados poderiam ser aprimorados com estas informações valiosas, aumentando cada vez mais a importância e a popularidade da mineração de dados nesta área.^[9,10] Métodos analíticos, incluindo algoritmos estatísticos tradicionais e de Inteligência Computacional (IC), são usados para melhorar a eficiência do ciclo científico. Existe uma integração colaborativa de métodos estatísticos tradicionais e de mineração de processos e de dados com o intuito de exploração das informações dos registos eletrónicos de saúde (RES) para estudar grupos de doentes em várias aplicações clínicas e de pesquisa, como extração de fenótipos, medicina de precisão, avaliação de intervenções, previsão, deteção e progressão de doenças.^[9,11] Os algoritmos de IC são poderosas ferramentas de classificação que, quando utilizadas neste campo, podem criar modelos úteis para os profissionais de saúde para alocarem os seus recursos e focar-se de forma mais eficaz no tratamento dos doentes.^[12,13]

Apesar do principal objetivo dos RES ser a melhoria da gestão das informações de saúde dos doentes estes registos têm sido amplamente utilizados para um propósito secundário na investigação clínica e no aperfeiçoamento da prática médica.^[14] Este uso secundário de grandes conjuntos de dados pode ser uma fonte-chave de informações para redesenhar serviços e adaptar intervenções distintas às necessidades individuais dos pacientes.^[15] Fornecem um conjunto detalhado de informações que incluem dados demográficos, histórico médico, medicação, resultados de análises laboratoriais e diagnósticos. O investigador deve reconhecer que estes dados, não coletados

primariamente para fins de pesquisa, requerem limpeza e transformação antes de serem utilizados para análise. A natureza variada das necessidades médicas dos pacientes ao visitarem um hospital gera várias características nos dados, tornando a mineração de dados um desafio complexo. Essas características incluem a temporalidade, irregularidade, desequilíbrio de dados e limitações de esparsidade, juntamente com preocupações éticas associadas a utilização dos registros.^[11]

Para além dos RES existem outras áreas em que o enquadramento único de saúde com base em metodologias de comunicação inovadoras tem fornecido um grande volume de dados que requerem engenheiros e cientistas de dados com especialização em mineração de dados, IC e análise de dados.^[16] A importância da Internet of things (IoT) nesta área esta a aumentar e com o auxílio da rede 5G é possível capturar informações valiosas a partir de várias fontes.^[16,17] Os sistemas ciberfísicos e a IoT estão a ser amplamente utilizados para substituir a estrutura de comunicação existente. Este esforço visa melhorar a operação, reduzir custos e aprimorar a qualidade. No contexto dos cuidados de saúde, sistemas inteligentes de IoT têm capacidade para recolher, agrupar e transmitir uma vasta gama de fontes de dados.^[18] Estas informações podem ser classificadas em modelagem interpretável e modelagem preditiva, utilizando técnicas de mineração de dados e recurso a algoritmos de IC. No ambiente clínico, parâmetros vitais de saúde como eletrocardiograma, saturação de oxigênio, frequência cardíaca, frequência respiratória e pressão arterial são utilizados para fornecer cuidados preventivos a doentes com patologias cardiovasculares, renais, distúrbios neurológicos e cerebrovasculares.^[19]

A IC é então uma abordagem valiosa na ciência de dados e exerce uma crescente influência na pesquisa médica, com um ritmo acelerado de inovação. Este desenvolvimento é impulsionado, em parte, pela enorme expansão na capacidade computacional e disponibilidade de dados.^[12] À medida que a necessidade de políticas baseadas em evidências se torna cada vez mais urgente, os grandes dados e as ferramentas analíticas associadas prometem estratégias promissoras para promover políticas de saúde pública.^[15] Pode-se considerar o exemplo da capacidade de adaptabilidade e a potencialidade na gestão da crise de saúde gerada pela COVID-19. Conclui-se que com estas ferramentas é possível escalar, e as técnicas baseadas em dados processam informações provenientes de todo o mundo. Possibilitando aplicar estas técnicas como modelação de transmissão, previsão de disseminação, detecção de surtos e identificação precoce de casos.^[20] Desenvolvendo assim a capacidade de analisar as interações complexas e dinâmicas entre comportamento humano, genética e o ambiente.^[15]

A predição pode reduzir substancialmente os custos de saúde ao identificar pessoas mais propensas a doenças crônicas e ao facilitar intervenções precoces antes que problemas afetem a população como um todo, enquanto a inferência científica frequentemente exerce maior influência na compreensão mecanicista.^[9,12] O desenvolvimento de perfis populacionais mais precisos oferece a oportunidade de desenvolver novas abordagens para implementar estratégias de prevenção e a estratificação de risco melhorando a forma como os recursos são alocados (por áreas geográficas, grupos populacionais, doenças e outros desafios de saúde).^[1] O investigador tem que ter em conta que a escolha

de uma medida adequada para o contexto é vital, uma vez que a precisão em uma dessas medidas pode não o ser em outros contextos assim como pode não estar relacionada com uma medida clinicamente significativa. Em contraste a utilização de valores P, intervalos de confiança e intervalos credíveis dos alvos inferenciais e estimativas para estatísticas populacionais tendem a ser menos ambíguos, e a incerteza pode ser mais claramente caracterizada.^[12]

No âmbito dos cuidados de saúde existe uma resistência a mudanças tecnológicas em larga escala, atualmente as principais técnicas usadas nesta área incluem classificação, regressão e agrupamento. Os métodos de mineração de dados utilizados em diferentes técnicas dependem das características dos dados e dos contextos onde estão inseridos.^[9] De facto, estas técnicas têm apresentado resultados promissores em diversas tarefas, como rastreamento da progressão ou trajetória de uma doença, identificação de coortes, previsão de riscos à saúde e deteção de eventos adversos, diagnósticos, prognósticos, classificações, na criação de modelos preditivos e análise de fatores de risco de várias doenças.^[10,11] Considere-se os métodos mais reconhecidos por “caixa preta”, poderosos na sua capacidade de aprendizagem, todavia, com uma opacidade que dificulta a interpretação direta dos resultados obtidos: Floresta aleatória - altamente flexível, amplamente adotada; Redes neuronais- identifica relações complexas de forma implícita, mas a lógica subjacente aos resultados não é facilmente compreendida; Máquina de Vetores de Suporte- utiliza algoritmos de classificação para tarefas de duas categorias. Como métodos mais transparentes e interpretáveis pode-se referir: Regressão Logística- o valor dos coeficientes reflete a sua

importância relativa na influência do resultado da classificação; *Naive Bayes*- eficiente mesmo com pequenos conjuntos de dados, mas depende da suposição de independência condicional entre características, o que pode limitar o seu desempenho em cenários onde essa suposição não é válida; Árvore de Decisão- pode trabalhar com dados incompletos ou com erros durante o treino, requer grandes conjuntos de dados para produzir resultados precisos, mas oferece uma abordagem visualmente intuitiva para a tomada de decisões; Mineração de processos- destaca-se pela transparência mas a compreensão do domínio é crucial para extrair insights significativos, sendo menos opaco na exposição dos padrões e processos subjacentes. ^[11]

Mas tudo depende principalmente da qualidade e do tipo de dados que são utilizados. Antes do desenvolvimento de qualquer modelo baseado em dados, é crucial abordar desafios, limitações e problemas de qualidade nos RES. Estes dados tendem a ser heterogêneos, representando informações estáticas, por exemplo informações demográficas ou podem conter informações temporais que são adquiridas ao longo de múltiplas visitas. Exemplos dos dados estruturados (numéricos ou categóricos) incluem códigos médicos, dados administrativos e sinais vitais. A qualidade de algumas características demográficas (como idade e gênero) frequentemente é razoável devido à simplicidade no registo dessas características, enquanto também têm uma relação significativa com condições de saúde. Outras características, como estado civil ou situação socioeconómica, podem não ser registradas, já que frequentemente são marcadas como campos opcionais em alguns sistemas de recolha de dados. Algumas pesquisas consideram altura, peso, índice de massa corporal (IMC) e circunferência da cintura como

sinais vitais, uma vez que são indicadores significativos de saúde e bem-estar. Relativo a códigos médicos que padronizam a representação da informação, um exemplo para representar diagnósticos médicos e procedimentos é a Classificação Internacional de Doenças (ICD). Trata-se de uma ferramenta globalmente utilizada para epidemiologia, gestão de saúde e fins clínicos, que padroniza a representação de doenças, transtornos, lesões e outras condições de saúde relacionadas.^[11]

Conjuntos de dados independentemente de seu tamanho, estão sujeitos a vieses que precisam ser bem compreendidos e considerados para evitar interpretações errôneas e conclusões incorretas.^[1] A generalização da informação registrada em instituições de saúde pode levar à perda de detalhes específicos. Em estudos baseados em RES, critérios como restrições de idade, informações clínicas ausentes ou limitadas, e presença/ausência de condições médicas afetam a validade e a aplicabilidade da análise a uma população mais ampla (chamado de viés de seleção). A temporalidade é inerente, com a variação no número e intervalo de visitas dos pacientes dificultando a comparação entre eles. Pacientes com maior necessidade médica frequentemente têm mais visitas, o que pode gerar viés nos dados e impactar os resultados da análise. Assim estes dados dos registros são irregulares e apresentam desequilíbrio de classes, onde a representação de casos de doenças graves é muito menor em comparação com casos menos graves ou normais. Além disso, são caracterizados por uma natureza esparsa, com muitos valores em falta devido à variabilidade nas necessidades médicas individuais e à diversidade das informações registradas em cada visita para pacientes distintos.^[14]

No centro da análise e interpretação destes conjuntos de dados devem estar padrões metodológicos sólidos.^[1] O foco está na reprodutibilidade dos resultados. Portanto, é necessário concentrarmo-nos inicialmente na qualidade dos dados e nos cenários de completude dos mesmos. A confiança na robustez e estabilidade das análises e relatórios é vital, tornando-se crucial a engenharia de dados, bem como a escolha do modelo. Por exemplo os investigadores que desenvolveram trabalhos com dados da pandemia COVID-19, utilizaram várias técnicas para prever muitos aspetos da doença, incluindo a taxa de mortalidade e a gravidade. Poderá se considerar complexa a reprodutibilidade desta pesquisa devido às metodologias utilizadas para recolher dados, à qualidade dos dados, à descrição da abordagem metodológica para treinar modelos de previsão, à sobre utilização de imputação de dados e ao sobre ajustamento. Os investigadores em saúde devem incluir nos seus estudos a divulgação de todo o código e fornecer declarações claras sobre o ajuste do modelo e os dados retidos utilizados para reportar a precisão, de modo a facilitar a avaliação externa da reprodutibilidade dos resultados.^[12,13] Uma revisão sistemática da literatura sobre modelos de previsão para diagnóstico e prognóstico da COVID-19 relata que a maioria dos modelos apresenta um alto ou um risco não claro de viés, principalmente devido ao sobre ajustamento do modelo, avaliação inadequada do modelo e tratamento inadequado de dados em falta.^[13]

Um grande objetivo da mineração de dados em RES deve ser contribuir para o benefício público, melhorando o sistema de saúde. Assim assegurar a equidade nesse processo é fundamental, evitando que características sensíveis, como género, idade, raça e orientação sexual,

distorçam as decisões no domínio da saúde.^[14] Mesmo que um sistema tenha um bom desempenho geral, pode falhar em regiões sub-representadas, muitas vezes representando minorias na sociedade, afetando a precisão das decisões críticas. Dados sociais, frequentemente não representam adequadamente minorias. O Viés de Representação nos dados pode surgir por várias razões, desde discriminação histórica até problemas na recolha e preparação dos dados. Pode ser causado por viés de seleção, quando a amostragem atinge apenas uma parte da população ou quando a população de interesse difere da utilizada durante o treino do modelo. Esse viés não é novo e tem sido reconhecido nas áreas de mineração de dados, gestão de bases de dados e estatística. No contexto da ciência de dados responsável e da IA confiável, houve um foco renovado no problema, com abordagens inovadoras.^[21] Frequentemente os atributos demográficos não estão disponíveis, o que dificulta a avaliação da equidade dos classificadores. Deverá ser aprofundada a relação entre desempenho de classificação e quantificação, explorando a possibilidade de separar esses objetivos. O foco na equidade algorítmica visa abordar o viés nos conjuntos de dados, independentemente de como os dados são utilizados. Existem três perspetivas de equidade: individual, de grupo e de subgrupo. Nos sistemas de decisão automatizados, a representação e diversidade nos conjuntos de dados, especialmente demográfica, são essenciais para garantir resultados justos e precisos. Para enfrentar o viés de representação, pode-se considerar adicionar mais dados ou, quando isso não é viável, procurar soluções preventivas, como informar os usuários sobre o problema ou ajustar consultas para atender às restrições de representação. É crucial envolver especialistas do domínio para avaliar e validar semanticamente os grupos/regiões

identificados, pois nem todos os subgrupos sub-representados são significativos ou válidos.^[21, 22] Uma potencial abordagem é orientar os algoritmos para maximizar o desempenho de quantificação, mantendo baixo o desempenho de classificação, garantindo, assim, a privacidade individual e estimativas precisas em nível de grupo. Além disso, é importante explorar intervalos de confiança para estimativas de equidade geradas por métodos de quantificação. Isso oferece uma indicação confiável da confiança nas estimativas de equidade de grupo, algo crucial para investigadores que procuram recursos e atenção para os efeitos de modelos em diferentes populações.^[22]

Também a gestão de valores em falta, especialmente em dados sensíveis como os da área da saúde, pode originar resultados falsos positivos e alterar as percepções. Muitas vezes, os dados (especialmente os grandes conjuntos de dados) apresentam muitos valores em falta, afetando a validade de qualquer investigação e análise. A análise de dados com valores em falta reduz a precisão. Por outro lado, ignorar todos os valores em falta durante a investigação conduz a análises ou resultados enviesados.^[23]

Por fim de referir, apesar de não ser o alvo do trabalho, a alta sensibilidade dos dados de saúde, trazendo consigo uma potenciação ética, legal e de compliance gerando um alto risco de divulgação e exposição de informações. À medida que os sistemas de saúde se tornam inteligentes e ubíquos, é essencial proteger a sua segurança e a confidencialidade dos dados, a cibersegurança na rede de comunicação não pode ser ignorada.^[10,24]

2.2 Estudos nacionais com foco em DALYs

Nesta segunda revisão procuramos o estado da arte neste âmbito específico e a nível nacional, assim surgiu a seguinte query: (Epidemiologic Methods [MeSH] OR Data Collection [MeSH] OR Vital Statistics [MeSH]) AND (Life Expectancy [MeSH] OR Disability-Adjusted Life Years [MeSH] OR Cost of Illness [MeSH]) AND Portugal [MeSH].

A pesquisa foi realizada na PUBMED, com os seguintes critérios:

- de inclusão - artigos em língua portuguesa ou inglesa, publicados entre 2022-2023;
- de exclusão - documentos de conferencias, artigos de opinião, cartas ao editor.

Daí resultaram 6 artigos que fornecem perspectivas exploradas nesta tese em que se destacam a importância da metodologia sólida na análise dos impactos na saúde e reforçam a importância da formulação de políticas de saúde. São abordados temas como cancro renal, poluição do ar, doenças cardiovasculares, envelhecimento populacional e varicela, todos enfatizam a necessidade de métodos precisos para compreender melhor os impactos na saúde.

O estudo relacionado com o cancro renal foca-se na quantificação dos anos de vida perdidos, anos de vida produtiva perdidos e a avaliação financeira presente da perda de produtividade futura devido a mortalidade precoce em Portugal e na Grécia. Fornece insights para orientar políticas de saúde e destacando a urgência de implementar estratégias mais eficazes de prevenção e tratamento. ^[24]

O estudo referente à poluição do ar em Portugal apresenta uma análise detalhada sobre os efeitos com foco na associação entre a presença de poluentes como PM2.5, PM10, NO2 e O3 e as taxas de mortalidade. Salienta a necessidade premente de revisão metodológica, enfatizando a importância da seleção apropriada de marcadores para prever com precisão os impactos na saúde. É explicado as funções usadas no modelo para responder às concentrações de poluentes. Isso inclui a consideração de riscos relativos e taxas de incidência base para estimar proporções atribuíveis, incidências excessivas e casos associados à exposição a esses poluentes. Não se restringe apenas a observar mudanças nos níveis de poluentes ao longo do tempo, mas também procura correlacionar essas variações com diversas fontes potenciais, como atividades industriais, transporte, eventos naturais como incêndios florestais e mudanças econômicas. Destaca-se, ainda, os impactos na saúde que emergem dessas variações.

Para a análise, são utilizados métodos estatísticos complexos, como os Modelos Lineares Mistos, que consideram correlações e diferenças nos dados coletados ao longo do tempo e em diferentes municípios. O estudo explora várias estruturas de variância-covariância para ajustar os dados e utiliza critérios estatísticos para comparar e selecionar modelos, com o objetivo de oferecer resultados robustos e confiáveis.^[25]

Relativo à carga de doenças cardiovasculares o artigo aborda as limitações metodológicas inerentes ao Global Burden of Disease Study (GBD) ao analisar a carga de doenças cardiovasculares nos países de língua portuguesa. Destaca as dificuldades causadas pela heterogeneidade dos dados nos territórios com informações de qualidade inferior.

São reveladas limitações na modelagem da coexistência de vários fatores de risco simultâneos e a possibilidade de subestimar o risco quando múltiplos

fatores estão presentes. Além disso, ressalta a falha em abordar adequadamente influências socioculturais, demográficas e étnicas, afetando a precisão da captação desses elementos nos modelos analíticos.

Apesar dessas limitações, o GBD é reconhecido como uma metodologia robusta e válida para estimar a carga de doença cardiovascular atribuível aos fatores de risco. Mesmo na ausência de dados primários, fornece métricas comparáveis entre os países, oferecendo insights relevantes para a formulação de políticas de saúde.^[26]

Através do artigo que aborda o EpiDoC (uma coorte populacional nacional em todo o país) é possível a análise detalhada sobre o envelhecimento da população idosa em Portugal ao longo de 10 anos, focando-se na qualidade de vida relacionada à saúde e na função física. Os dados revelam um aumento na percentagem de idosos, assim como um declínio na saúde percebida e na capacidade física, especialmente entre as mulheres e à medida que a idade avança. Além disso, fatores como educação, exercício físico regular, multimorbidade e peso excessivo influenciam significativamente essas trajetórias de saúde. Podemos concluir a necessidade de políticas públicas mais informadas e recursos direcionados para cuidados de longo prazo e saúde, com foco na promoção da saúde e na redução das disparidades entre a expectativa de vida e a expectativa de vida saudável.^[27]

No artigo sobre a varicela e a sua prevalência conclui-se que esta sofre de subnotificação e subavaliação, dificultando a compreensão real do impacto da doença. Por isso é crucial avaliar esse impacto na qualidade de vida para direcionar recursos adequadamente.^[28]

De uma forma transversal os artigos abordam a necessidade contínua de aprimorar os métodos de pesquisa para direcionar políticas de saúde pública

mais precisa e eficaz. Assim investir em métodos robustos é essencial para compreender melhor os impactos na saúde e promover o bem-estar da sociedade.

3. Metodologia

A população alvo deste estudo são os utentes adultos (≥ 18 anos) que recorreram, pelo menos uma vez, a qualquer consulta médica (programada ou não programada) no Grupo Lusíadas Saúde (Braga, Porto, Lisboa, Almada, Amadora, Sacavém, Vila Nova de Gaia, Albufeira e Faro) num período de 12 meses (2022- 2023).

Foi um estudo observacional transversal baseado numa base de dados de registos eletrónicos médicos do grupo Lusíadas Saúde. Este grupo privado de saúde está representado em diferentes regiões nacionais (Braga, Porto, Lisboa, Almada, Amadora, Sacavém, Vila Nova de Gaia, Albufeira e Faro) por hospitais e clínicas que fornecem serviços de assistência clínica, intervenções cirúrgicas, atendimento permanente, internamento, serviços de imagiologia e de exames especiais. Os clientes procuram os seus serviços de uma forma integral ou em suporte ao serviço nacional de saúde.

A recolha de dados possibilitou a análise e caracterização da população que recorreu pelo menos uma vez a consulta programada ou não programada, entre 1 de agosto 2022, até 1 agosto 2023, ao Grupo de Saúde Lusíadas. As variáveis recolhidas foram retiradas do Processo Clínico Eletrónico, do último registo disponível, registado pelos médicos dos hospitais e clínicas: patologia na terminologia ICD9, sexo, estado civil, idade, escolaridade, código postal do concelho residência, peso, altura.

Após parecer positivo do comité de ética do Grupo Lusíadas Saúde e do Encarregado da proteção de dados, os dados foram extraídos do servidor, pelo Departamento operações e sistemas de informação. Foi utilizada uma plataforma de processamento e edição de dados anonimizados, entregues com segurança ao investigador, garantindo todos os requisitos do RGPD.

Para a análise dos dados, foi fundamental realizar um criterioso pré-processamento para verificar a integridade dos dados, identificar e tratar valores ausentes ou inconsistentes, converter formatos de dados, e avaliar a necessidade de codificar ou agrupar categorias para simplificar a análise.

Utilizou-se a linguagem R para a limpeza e transformação dos dados. Esta fase envolveu a verificação de valores nulos ou duplicados, a imputação de dados ausentes e a padronização de variáveis categóricas. O pacote ``dplyr`` foi especialmente útil para manipulação de dados, enquanto que ``tidyr`` auxiliou na organização e limpeza dos dados. Foram aplicadas técnicas de transformação e normalização.

Este estudo apresenta uma análise descritiva que caracteriza a população em estudo. Foram utilizadas várias técnicas estatísticas para descrever as características biodemográficas e o perfil de morbilidades da população. Ferramentas como ``ggplot2`` permitiram a visualização de dados de forma clara, facilitando a compreensão de padrões e tendências dentro do conjunto de dados.

A seguir, foi aplicado o algoritmo de clustering Kmeans. Este método foi escolhido devido à sua eficiência na segmentação de grandes conjuntos de dados e à capacidade de identificar grupos naturais dentro da população. Utilizou-se a função ``kmeans`` do R para agrupar os indivíduos com base em características semelhantes. A determinação do número adequado de clusters foi feita através do método do cotovelo e da análise do índice de

silhueta, garantindo assim a formação de clusters significativos. Os resultados obtidos revelaram padrões e grupos distintos dentro da população, fornecendo informações que possibilitaram o desenvolvimento de um plano de ação mais direcionado.

4. Resultados e discussão

Os dados foram disponibilizados sob a forma de três ficheiros CSV, representando cada um uma área geográfica das unidades clínicas. Base de dados Norte inclui Porto, Gaia e Braga; a do Centro inclui Lisboa, Amadora, Sacavém e Almada e a do Sul que inclui Albufeira e Faro.

As variáveis incluídas são: identificador aleatório do doente (DOENTE), sexo (SEXO), nível de escolaridade (ESCOLARIDADE), estado civil (ESTADO_CIVIL), código postal com quatro dígitos (COD_POSTAL), data de nascimento do doente (DT_NASC), altura (HEIGHT), peso (WEIGHT), todos os códigos ICD-9 associados ao doente (COD_ICD9) e descritivos dos códigos ICD-9 (DSC_ICD9).

Para mitigar o risco de privacidade dos dados foram tomadas medidas para garantir a proteção de informações sensíveis. Assim para redução do risco de identificação pela data de nascimento foi calculado a idade dos doentes tendo como referência “2023-12-31” e eliminadas todas as colunas de data de nascimento. Dado que pretendíamos estudar a população adulta foram apagados todos os registos de doentes abaixo dos 18 anos.

R é uma escolha amplamente utilizada entre profissionais de ciência de dados e investigação dada a sua facilidade de uso e recursos disponibilizados. A sua sintaxe simples facilita a aprendizagem e aplicação de análises estatísticas complexas, mesmo para aqueles sem um forte background em

programação. Desenvolvida especificamente para análise estatística e gráficos, a linguagem R beneficia de uma extensa documentação e de uma vasta comunidade de utilizadores, proporcionando tutoriais, exemplos e suporte facilmente acessíveis, o que acelera o processo de aprendizagem e aplicação. Além disso, o R oferece acesso a uma vasta gama de pacotes gratuitos através do Comprehensive R Archive Network (CRAN), que cobre uma ampla variedade de técnicas estatísticas, machine learning, visualização e manipulação de dados. Estes pacotes, desenvolvidos e mantidos por uma comunidade ativa de investigadores, garantem ferramentas sempre atualizadas com as últimas metodologias e tecnologias. A longa história de utilização do R em investigação académica e científica contribui para a sua credibilidade e aceitação, refletindo a sua robustez e precisão. Como software de código aberto, o R promove a transparência nas análises, permitindo que outros investigadores verifiquem e reproduzam estudos com facilidade. Assim foi utilizada esta ferramenta em todo o trabalho. Foi realizada uma análise detalhada da qualidade dos dados e aplicadas técnicas de pré-processamento, como limpeza e normalização dos dados. As variáveis SEXO, COD_POSTAL, COD_ICD9 DSC_ICD9 foram transformadas em fatores e os *missings* em NA.

As características dos dados retirados de RES têm implicações significativas na análise e interpretação dos resultados em pesquisa de saúde e epidemiologia. A compreensão e o uso adequado dessas características são essenciais para realizar análises robustas e obter conclusões válidas. Assim a temporalidade refere-se à dimensão do tempo nos dados, neste caso os dados foram recolhidos durante um ano. O desequilíbrio de dados ocorre quando algumas classes ou categorias de dados são muito mais representadas do que outras, sendo que neste caso temos uma população

muito específica que recorreu a um grupo de saúde durante esse período de tempo reconhecendo-se a impossibilidade de qualquer validação para além deste grupo ou formulação de qualquer generalização para a população portuguesa. Por último uma menção à esparsidade que se refere aos valores ausentes e que acabou por condicionar a caracterização da população não permitindo ser mais completa. Também aqui, “o retirar frutos” destes dados, poderiam trazer um impacto positivo para que os médicos completassem os registos mesmo os que não são de registo obrigatório.

4.1 Caracterização da população em estudo

Numa primeira análise dos descritivos de ICD9 deparamo-nos com códigos associados a exames médicos e procedimentos que não foram acautelados para serem excluídos no pedido da base de dados. Assim foi realizada, baseada na tabela do ICD9 publicada online pela OMS uma filtragem e retirados o grupo de códigos E e V - causas externas de lesões e classificação suplementar.

Como será apresentado nas próximas tabelas existem variáveis com uma alta taxa de missings assim foi decidido com recurso ao estudo de bibliografia retirar-las da análise posterior à caracterização. Porque quando uma variável tem uma taxa de missings acima de 70%, a validade e a precisão das análises podem ser significativamente comprometidas. A remoção desses dados evita o viés e melhora a robustez das análises. É possível realizar imputação dos valores para dados, mas em grandes proporções pode introduzir uma variabilidade não representativa da realidade, tornando as imputações pouco confiáveis. Métodos de imputação, como o Multiple Imputation, são

mais eficazes quando a taxa de missings é baixa. Altas taxas de missings dificultam a obtenção de imputações plausíveis e podem resultar em maior incerteza nos dados. Assim em análises estatísticas e de ciência de dados, a prática comum é remover variáveis com uma alta taxa de missings para manter a integridade dos dados e evitar análises enviesadas. Em grandes bases de dados, como em RES é comum excluir variáveis com muitos missings para simplificar a análise e evitar a introdução de vieses complexos. Assim foram retiradas as variáveis com mais de 70% de missings pela necessidade de manter a qualidade e a integridade dos dados, evitando vieses e incertezas nas análises. [29,30,31]

A tabela seguinte explana a caracterização geral da população adulta utilizadora (pelo menos 1 vez ao longo de 12 meses) das unidades Lusíadas Saúde em Portugal Continental.

Tabela 2 Descrição das bases de dados – considerando adultos (≥ 18 anos)

Norte, n = 131380

Sexo, n(%)		Missings, n(%)
Feminino	82459 (62.7)	706(0.5)
Masculino	48224(36.7)	
Idade (anos), M$\pm$sd, min-máx	51.4 $\pm$ 18.4, 18-124	0
Peso (kg), M$\pm$sd, min-máx	73.3 $\pm$ 43.2, 0-992	119939(91.3)
Altura	6.92 $\pm$ 32.3, 0-1671	119598(91)
Escolaridade		131339(99.9)
Estado civil		108715(82.7)
Código ICD9		0
Descritivo ICD9		0
Código Postal		3419(2.6)

Centro, n = 173709

Sexo, n(%)		Missings, n(%)
Feminino	110644 (63.6)	146(0.1)
Masculino	62919(36.2)	
Idade (anos), M$\pm$sd, min-máx	51.1 $\pm$ 18.6, 18-124	0
Peso (kg), M$\pm$sd, min-máx	72.6 $\pm$ 37.1, 0-960	150869(86.9)
Altura (cm)	4.6 $\pm$ 22, 0.11-200	150556(86.7)
Escolaridade		173626(99.9)
Estado civil		134893(77.7)
Código ICD9		0
Descritivo ICD9		0
Código Postal		4421(2.5)

Sul, n = 52611

Sexo, n(%)		Missings, n(%)
Feminino	29793 (56.6)	542(1)
Masculino	22276(42.3)	
Idade (anos), M$\pm$sd, min-máx	51.6 $\pm$ 18.7, 18-124	1(0.002)
Peso (kg), M$\pm$sd, min-máx	73.8 $\pm$ 20.2, 0-584	50931(96.8)
Altura (cm)	7.3 $\pm$ 29.7, 0-182	50868(96.6)
Escolaridade		52599(99.9)
Estado civil		42429(80.6)
Código ICD9		1(0.002)
Descritivo ICD9		1(0.002)
Código Postal		4762(9.1)

Figura 1 Gráfico do Top 10 ICD9 na base de dados Norte

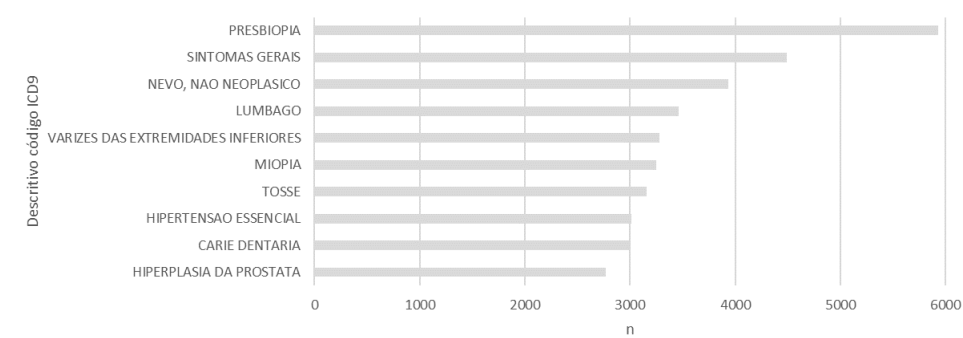


Figura 2 Gráfico do Top 10 ICD9 na base de dados Centro

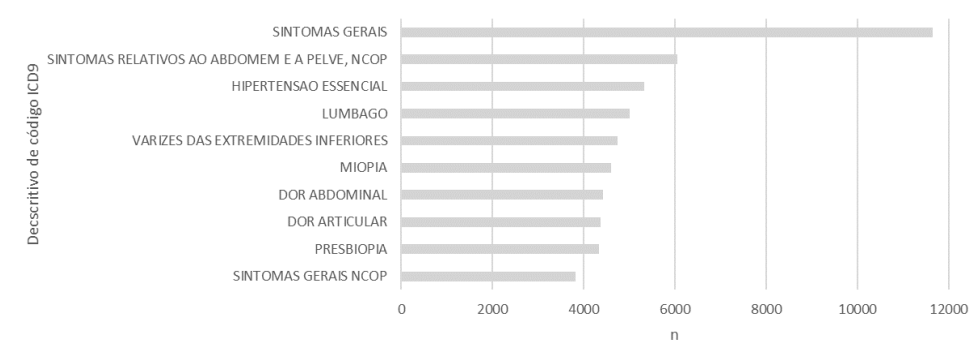
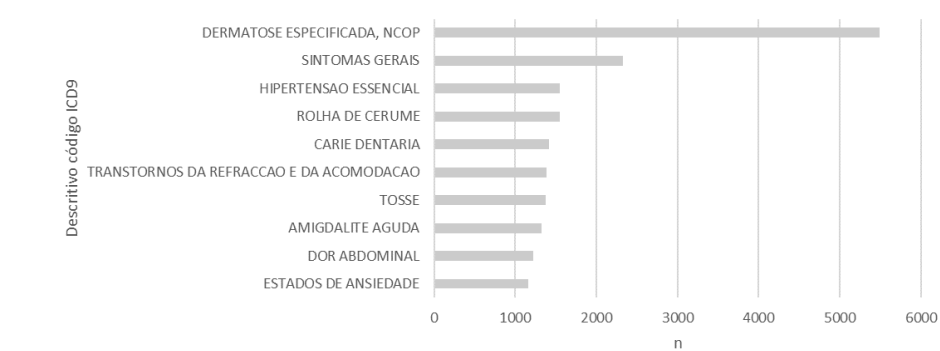


Figura 3 Gráfico do Top 10 ICD9 na base de dados Sul



Analisando estes gráficos com o top 10 de descritivo de códigos ICD9 não existe a informação necessária para delinear intervenções nos moldes a que nos propusemos. Foram inseridos diagnósticos generalistas como sintomas

gerais, de sinais como tosse ou alguns de baixa relevância clínica como rolha de cerume. De salientar a presença neste Top 10 da patologia relacionada com a taxa de DALYs nacional o Lumbago tanto na base de dados do Norte como do Centro. Neste ponto temos a caracterização possível da população presente na base de dados e a informação dos Top 10 de cada uma das bases de dados.

4.2 Caracterização da população em estudo após filtragem

Apesar de cumprido o primeiro objetivo do trabalho não é possível retirar informação clinicamente pertinente destes dados. Na continuação do apresentado na metodologia foi realizada uma filtragem nos códigos ICD9 de forma a considerarmos o Top 10 nacional de DALYs nesta base de dados.

Tabela 3 Correspondência DALYs – Código ICD9

Top DALYs Nacional	Código ICD9
Doença cerebrovascular	430 a 438
Doença isquémica do coração	410 a 414
Lumbago	724.2
Diabetes mellitus	250.00 – 250.93
DPOC	491, 492, 496
Depressão	296.2, 296.3, 311, 300.4
Tumor maligno do pulmão	162 – 162.9
Infeções respiratórias inferiores	466, 480-488, 490, 491, 492, 494
Tumor maligno do cólon e reto	153, 154
Doença Alzheimer	331.0

A tabela 3 apresenta as atribuições dadas de códigos ICD9 às Top 10 DALYS em Portugal, através de pesquisa bibliográfica de artigos do âmbito e das

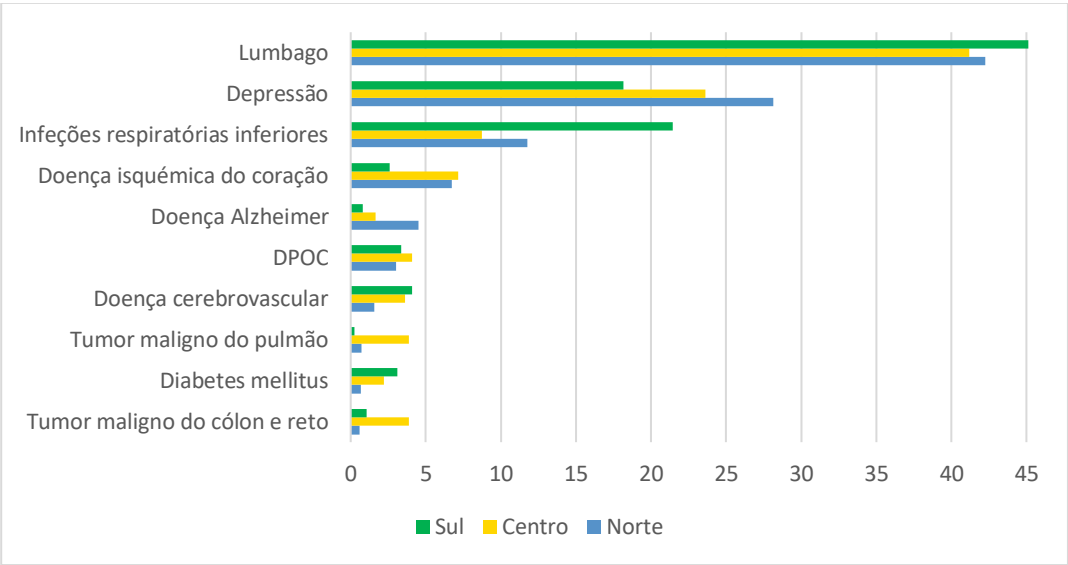
listagens de ICD9 publicadas pela OMS. Dado que o Grupo de Saúde selecionado apenas trabalhava com esta codificação esta poderia ser considerada uma limitação ao trabalho. Apesar desse aspecto existem estudos que sublinham a importância dos códigos ICD9 e demonstram a eficácia desses códigos na condução de pesquisas robustas. Há evidência que a aplicação dos códigos ICD9 em diferentes contextos de pesquisa se mostrou eficaz. Um estudo realizado pela Autoridade Regional de Saúde da Úmbria avaliou a precisão dos códigos ICD9 para acidentes vasculares cerebrais hemorrágicos e isquêmicos em bases de dados administrativas. Os resultados indicaram alta precisão diagnóstica dos códigos para hemorragia subaracnóidea, hemorragia intracerebral e outras hemorragias intracranianas não especificadas, com sensibilidade e especificidade entre 96% e 100%. Estes achados confirmam a fiabilidade destes códigos para estudos epidemiológicos e de serviços de saúde. Um outro sobre depressão apresenta variações na precisão dos códigos ICD9 e ICD10 em diferentes contextos. Uma revisão sistemática mostrou que códigos mais inclusivos tendem a ter altos valores preditivos positivos, mas sensibilidades subótimas. Para a Doença Pulmonar Obstrutiva Crônica (DPOC), os códigos ICD9 demonstraram alta precisão quando validados. A utilização desses códigos tem sido eficaz na descrição da epidemiologia da DPOC, avaliação da eficácia e segurança de tratamentos, e monitorização da qualidade dos cuidados. Foram considerados outros artigos que abordaram outras patologias e respetivos códigos ICD9 utilizados para codificar e categorizar diagnóstico como é o caso da diabetes, doenças cerebrovasculares e doenças isquêmicas do coração.^[32-38]

Com esta filtragem realizamos nova caracterização dos indivíduos que tenham pelo menos um diagnóstico coincidente com a lista TOP 10 de Portugal (de 2019) de DALYs.

Tabela 4 Descrição das bases de dados filtrado por DALYs – considerando adultos (>=18anos)

	Norte N = 9163		Centro N = 13490		Sul N = 2921	
Sexo, n(%)	Missings		Missings		Missings	
<i>Feminino</i>	5313	(57.9)	7655	(56.7)	1589	(54.3)
<i>Masculino</i>	3814	(41.6)	5817	(43.1)	1309	(44.8)
Idade (anos), M±sd, min-máx	59.5±18.1, 18-124	0	60±17.4, 18-104	0	60±17.1, 18-103	0
Peso (kgs), M±sd, min-máx	72.1±27.5, 1-719	7653 (83.5)	73.8±33.2, 1-863	10183 (75.5)	74.8±15.2, 0-128	2725 (93.2)
Altura (cm), M±sd, min-máx	129.2±68.5, 1-183	9135 (99.7)	131.8±69.7, 1-192	13417 (99.5)	73.6±81.4, 0-169	2908 (99.5)
Escolaridade		9159 (99.9)		13481 (99.9)		2920 (99.9)
Estado Civil		7094 (77.2)		9888 (75.3)		2170 (74.3)
Código ICD9		0		0		0
Descritivo ICD9		0		0		0
Código Postal		0		227 (1.7)		0

Figura 4 Gráfico das DALYs (%) nos doentes Lusíadas



As bases de dados apresentam o mesmo Top 3. Deparamo-nos com algumas taxas superiores no Centro para os Tumores malignos e na base de dados do Norte com maior evidencia da taxa de Doença de Alzheimer.

Comparando com as DALYs nacionais o top 3 obtido refere-se respetivamente às taxas que estão em 3º, 6º e 8º lugar nacional. Podemos extrapolar para o cariz de cuidados que o grupo representa atualmente, na sua maioria um conjunto de doentes com uma menor gravidade clínica imediata que recorre a estes serviços. Seria pertinente realizar no futuro uma nova abordagem de caracterização das principais patologias se nos próximos anos este tipo de instituições privadas adquirirem outro papel no sistema nacional de saúde. Este gráfico é então estritamente referente a uma população muito específica que recorreu aos cuidados das unidades Lusíadas. Assim com esta base de dados nunca poderemos generalizar para toda a população, mas não nos impede de obter informação relevante para os indivíduos que optaram por recorrer a esta instituição e que assim poderão ser devidamente acompanhados.

4.3 Oportunidades na intervenção clínica

Com foco na principal patologia identificada poderemos definir uma intervenção clínica nesta população e saber quais as áreas do país em que estes indivíduos residem. Assim nesta população existe um n para o lumbago acima de 400 no código postal 1500, seguido do código postal 8200 acima de 300 casos e de seguida o 1600 com um valor próximo dos 300. Todos os outros códigos são abaixo dos 200 casos. De uma forma segura e respeitando a privacidade poderemos utilizar georreferenciação dos códigos postais e

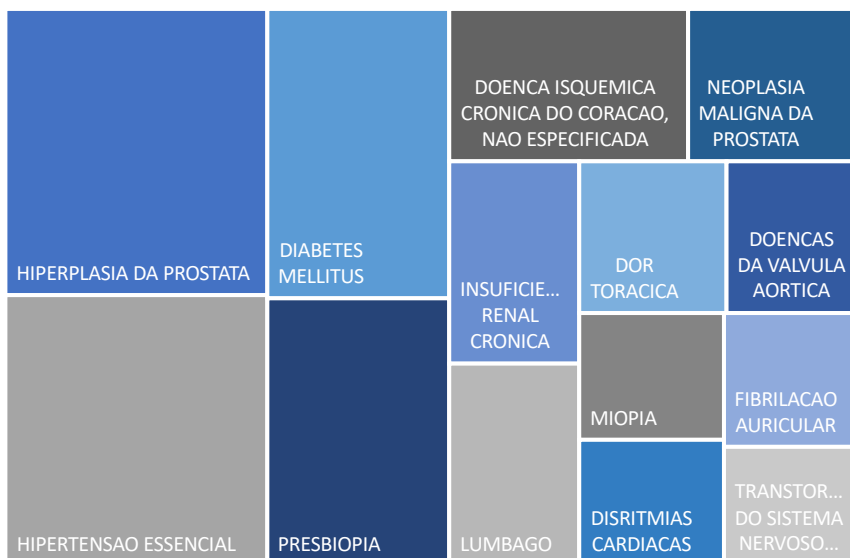
criar heatmaps interpretativos destes dados e outros no âmbito desta população para características biogeográficas, ambientais e recursos. Poderão ser desenvolvidos planos de ação e intervenções clínicas relevantes juntamente com as entidades locais.

Este acompanhamento e o estudo da investigação científica desta patologia pelas equipas poderá gerar o melhor conhecimento para a intervenção junto desta população. A prevalência de dor lombar crónica é impulsionada pelo envelhecimento, obesidade e maior disposição para relatar dor. Num estudo de 2014, a prevalência é maior em mulheres, idosos e áreas urbanizadas, com maior incidência na região Centro de Portugal. Indivíduos que praticam exercício físico regularmente apresentam menor prevalência de lombalgia. Outro estudo refere a necessidade de um acompanhamento intensivo para pacientes que não recuperam em 3 meses e reavaliações dentro das primeiras 12 semanas. Destaca-se a importância de desenvolver tratamentos mais eficazes e ferramentas específicas para avaliar o curso clínico da DL em diferentes faixas etárias. Recomenda-se um enquadramento biopsicossocial que inclua educação, autogestão, retoma de atividades habituais, exercício e medidas psicológicas, além de incentivar a atividade física, permanência no trabalho e uso prudente de medicação e cirurgia.^{39-41}

Por outro lado, também poderemos selecionar uma patologia que seja relevante clinicamente e trabalhar com essa população. Com as características e áreas para personalizar e adaptar intervenções para uma maior efetividade. Por exemplo selecionando doença isquémica do coração na base de dados do Norte, sabemos algumas características relevantes clinicamente dos doentes que recorrem ao grupo Lusíadas com esta patologia diagnosticada. A faixa etária com maior taxa, os principais códigos postais onde existe mais diagnósticos desta patologia e as comorbilidades

mais comuns associadas a estes doentes com doença isquémica do coração. No gráfico da figura 6, observamos as comorbilidades associadas a esta patologia. Se fosse este o alvo do trabalho das equipas poderíamos fazer um estudo bibliográfico da temática e encontrar a melhor discussão científica à luz de outros estudos e trabalhos em diferentes populações. Por exemplo, a análise abrangente das comorbilidades em dois artigos analisados revelou uma heterogeneidade significativa, mas incluía hipertensão, insuficiência cardíaca, fibrilação atrial, insuficiência renal crónica, diabetes mellitus e doenças metabólicas. Nestes artigos pacientes com estas comorbilidades apresentam prognósticos piores e um maior risco de eventos cardíacos subsequentes.^[42,43]

Figura 5 Gráfico com áreas das comorbilidades mais comuns da doença isquémica no grupo estudado



Todo este trabalho poderia ser desenvolvido estatisticamente para as áreas que fossem mais prementes para os profissionais de saúde intervir. Seria

importante seguir uma abordagem multidisciplinar da equipa de saúde incluindo dos sistemas de informação e gestão de dados no sentido de estabelecer objetivos e a criação de planos de trabalho.

4.4 Resultados com abordagem pela inteligência computacional

Foi realizada uma outra abordagem destas mesmas bases de dados. Propunha-se uma utilização de inteligência computacional com o intuito de obter informações relevantes e justificadas. Assim foi colocada como variável dicotómica as patologias DALYS.

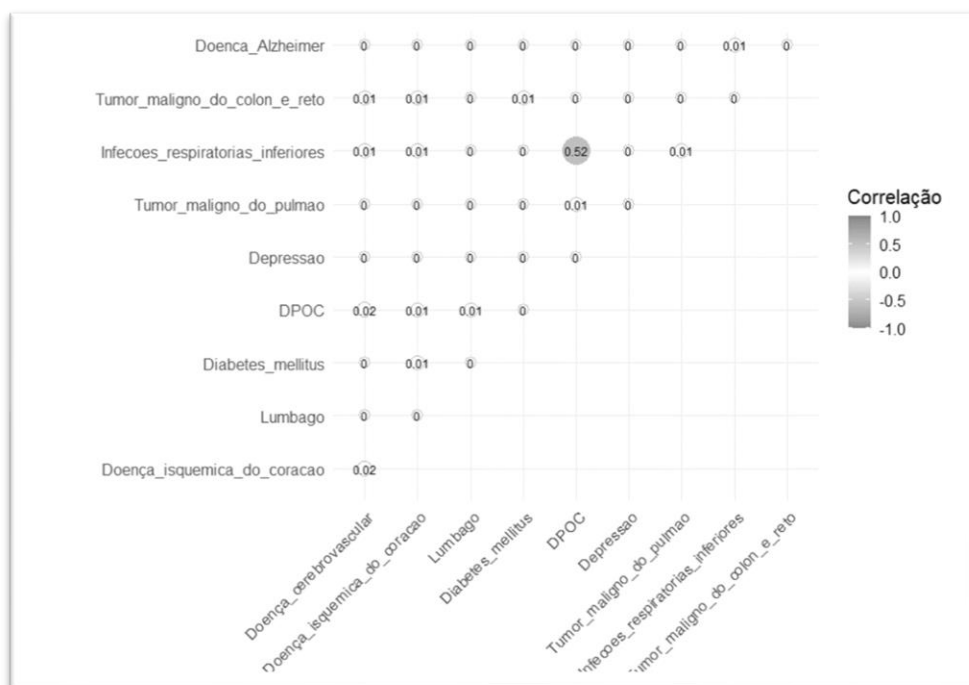
Figura 6 Exemplo das primeiras linhas da BD Norte com a variável dicotómica

DOENTE	SEXO	IDADE	COD_POSTAL	Doença_cerebrovascular	Doença_isquemica_do_coracao	Lumbago	Diabetes_mellitus	DPOC
3988167473	M	73	4435	0	1	0	0	0
3988183574	M	86	4490	0	0	0	0	0
3988198533	M	69	4050	0	0	0	0	0
3988346360	F	63	4430	0	0	0	0	0
3988398245	F	49	4420	0	0	0	0	0
3988413973	M	46	4250	0	0	0	0	0
3988436590	F	27	4465	0	0	0	0	0
3988482068	F	56	4490	0	0	0	0	0
3988524391	F	48	4455	0	0	0	0	0
3988551447	F	63	4715	0	0	0	0	0
Depressao	Tumor_maligno_do_pulmao	Infecoes_respiratorias_inferiores	Tumor_maligno_do_colon_e_reto	Doenca_Alzheimer				
0	0	0	0	0				
0	0	0	0	0				
0	0	0	0	0				
0	0	0	0	0				
0	0	0	0	0				
0	0	0	0	0				
0	0	0	0	0				
0	0	0	0	0				
0	0	1	0	0				
0	0	0	0	0				

Com a matriz de correlação apresentada, podemos fazer algumas observações sobre as relações entre as diferentes patologias no conjunto de dados. A maioria das correlações entre as patologias é bastante baixa, o que indica que, em geral, a presença de uma

condição de saúde não está fortemente associada com a presença de outra dentro deste conjunto de dados. Correlações próximas de zero indicam pouca ou nenhuma relação linear entre as variáveis. A mais alta presente na matriz é a correlação de 0.519 entre DPOC (Doença Pulmonar Obstrutiva Crônica) e Infecções Respiratórias Inferiores. Faz sentido clinicamente, porque doentes com DPOC são mais propensos a desenvolver infecções respiratórias devido à fragilidade do sistema respiratório.

Figura 7 Matriz de correlação das patologias na BD Norte



Outra forma de obter informação seria através da segmentação desta população em clusters, podendo personalizar as intervenções para atender às necessidades específicas de cada grupo. Recorrendo ao Kmeans criamos quatro clusters.

Tabela 5 Centros clusters

Doença Cerebrovascular	Doença isquêmica do coração	Lumbago	Diabetes mellitus	DPOC	Depressão	Tumor maligno do pulmão	Infeções respiratórias inferiores	Tumor maligno do colon e reto	Doença de Alzheimer
0,429	0,172	0,123	-0,024	21,819	-0,101	-0,022	11,334	-0,023	0,091
-0,001	-0,001	-0,001	-0,001	-0,045	0,001	-0,022	-0,024	-0,026	-0,001
-0,031	0,186	-0,158	-0,024	0,291	-0,129	44,996	0,615	-0,026	-0,051
0,442	0,659	-0,062	0,597	-0,046	-0,129	-0,02	-0,088	44,319	-0,051

Cluster 1: Alta Prevalência de Doenças Respiratórias: Alta prevalência de DPOC (21.818) e infecções respiratórias inferiores (11.334).

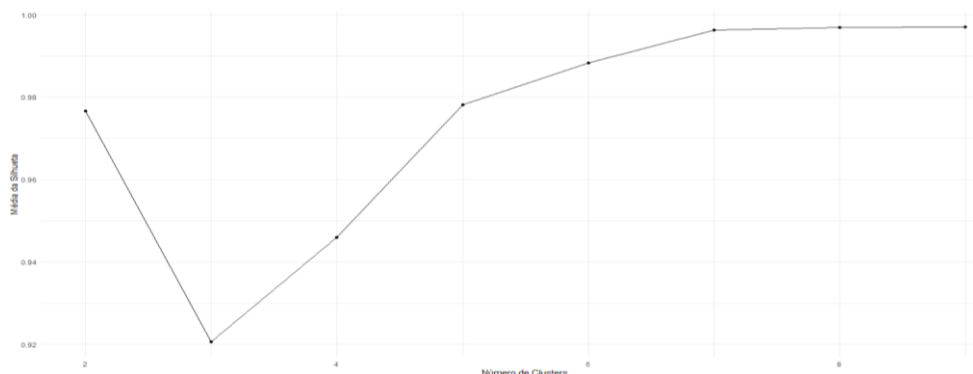
Cluster 2: Baixa Prevalência de Doenças: Valores médios próximos de zero ou negativos para todas as doenças, indicando uma baixa prevalência.

Cluster 3: Alta Prevalência de Tumores Malignos: Alta prevalência de tumores malignos do pulmão (44.995).

Cluster 4: Alta Prevalência de Doenças Cardiovasculares e Diabetes: Alta prevalência de doenças cerebrovasculares (0.441), doenças isquêmicas do coração (0.659) e diabetes mellitus (0.597). Tumores Malignos do Colon e Reto: Presença significativa de tumores malignos do colon e reto (44.319).

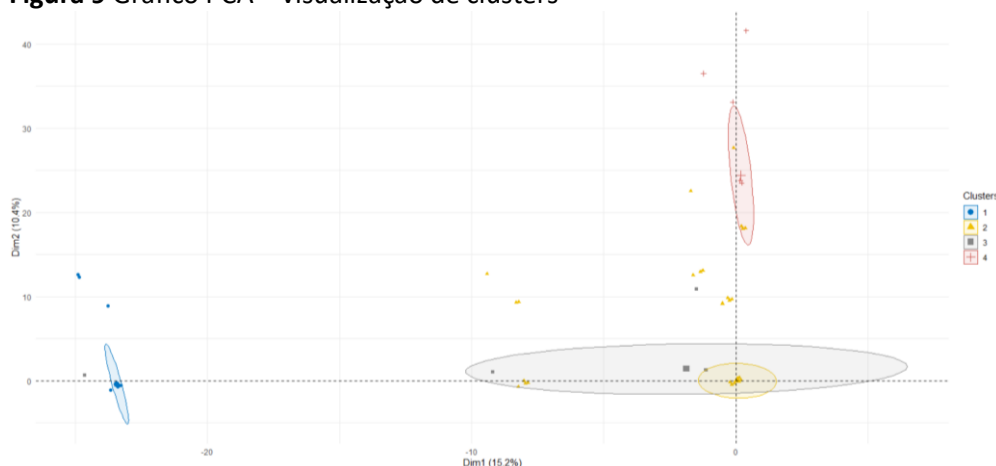
Para confirmação que esta seria a melhor abordagem dos clusters foram realizados alguns testes que comprovam como a melhor solução a criação destes 4 clusters.

Figura 8 Gráfico de análise silhueta



O gráfico de Análise de Silhueta mostra que para 3 clusters, a média da silhueta cai para cerca de 0.92, indicando uma qualidade inferior dos clusters. Para 4 clusters, a média da silhueta é significativamente maior (próxima de 0.96), indicando uma melhor definição e coesão dos clusters em comparação com 3 clusters. Além de 4 clusters, a média da silhueta continua a aumentar, mas a melhoria é menos acentuada, sugerindo que 4 clusters já fornecem uma boa qualidade de clustering.

Figura 9 Gráfico PCA – visualização de clusters



Com o Principal component analysis (PCA) confirmamos que 4 clusters é uma escolha adequada dada a separação clara entre os clusters e a coesão dentro de cada cluster.

Podemos a partir destas informações desenvolvidas estabelecer um plano de ação para realizarmos intervenções específicas e direcionadas. Focando-se nos principais fatores de risco comportamentais das doenças não transmissíveis.

Cada cluster deverá ter um gestor local que organiza a equipa multidisciplinar e realiza as pontes com as entidades locais necessárias. Poderão ser utilizados recursos locais disponíveis, como recintos desportivos, ginásios locais, grupos de

apoio comunitário e outras infraestruturas de saúde. Espera-se um prazo de 6 meses para implementação inicial, com revisões trimestrais.

A eficácia das intervenções será monitorizada através de indicadores de desempenho, como taxas de cessação do tabaco, redução do consumo de álcool, melhoria nos padrões alimentares, aumento da atividade física e gestão do IMC. Serão realizados levantamentos periódicos para avaliar a exposição aos fatores de risco e a adesão às recomendações. A análise dos dados permitirá ajustes nas estratégias de intervenção conforme necessário. A colaboração interdisciplinar será fundamental, integrando médicos, enfermeiros, nutricionistas, psicólogos e outros profissionais para garantir um suporte abrangente aos pacientes. Este plano visa reduzir a carga das DNTs na população, promovendo um impacto positivo e sustentável na saúde da comunidade.

No Cluster 1 foi identificada uma prevalência, nos doentes que recorreram ao grupo Lusíadas, de DPOC (Doença Pulmonar Obstrutiva Crônica) e Infecções respiratórias inferiores. Os códigos postais pertencem principalmente a freguesias próximas do centro do Porto (Lordelo do Ouro e Massarelos; Cedofeita, Santo Ildefonso, Sé, Miragaia, São Nicolau e Vitória, Ramalde; Paranhos). Existe também a considerar a freguesia de São Vicente em Braga. Intervenções propostas:

- ✓ estabelecer centros de reabilitação respiratória em ginásios locais para ajudar na recuperação e gestão da DPOC;
- ✓ campanhas de vacinação e educação: realizar campanhas de vacinação contra a gripe e pneumonia, juntamente com programas educacionais sobre prevenção de infeções respiratórias;
- ✓ grupos de apoio: criar grupos de apoio para pacientes com DPOC e suas famílias, oferecendo suporte e educação sobre a doença.

No Cluster 3 foi identificado uma prevalência nesta população específica, de Tumores malignos do pulmão. Mantemos o predomínio nas freguesias do Porto mas com a presença igualmente da freguesia da Madalena em Vila Nova de Gaia.

Intervenções propostas:

- ✓ programas de cessação do tabagismo;
- ✓ realizar programas educacionais sobre os riscos do tabagismo e a importância da deteção precoce do cancro;
- ✓ programas de apoio psicológico para doentes e familiares;
- ✓ atividades físicas adaptadas nos ginásios/piscinas e parques locais.

No Cluster 4 para os doentes que recorrem aos Lusíadas, durante o período em questão, podemos detetar uma prevalência de doenças cardiovasculares e diabetes e também de tumores malignos colon e reto. Surge uma distribuição mais abrangente por freguesias sendo no Porto (Lordelo do Ouro e Massarelos; Ramalde; Bonfim), em Gaia (Madalena; Vila Nova Gaia) e surgindo as freguesias de Leça da Palmeira e Senhora da Hora. Intervenções propostas:

- ✓ grupos comunitários/escolas locais sob orientação de profissionais de saúde para desenvolvimento de literacia em saúde - princípios fundamentais para um estilo de vida saudável;
- ✓ parcerias com ginásios locais e piscinas municipais para oferecer programas de exercícios físicos. Organização de caminhadas semanais em parques municipais.
- ✓ workshop alimentação saudável com recurso a mercados e feiras locais.

4.5 Discussão

No futuro, será um desafio enriquecedor utilizar outras bases que viabilizem estudos nacionais de prevalência e, eventualmente, apoiar o desenvolvimento de políticas públicas de saúde em Portugal. A existência de dados de elevada qualidade, obtidos a partir de registos clínicos, permitiria caracterizar a população de forma rigorosa, com vista à elaboração de planos de ação devidamente estruturados e fundamentados.

Adicionalmente, a automação da integração de RES em plataformas de investigação seria uma solução eficiente para a aquisição de informação clínica em tempo real. O processo de extração, transformação e carregamento pode ser facilitado através de ferramentas específicas ou *scripts* personalizados, que deveriam incluir procedimentos de pseudoanonimização dos dados, substituindo identificadores diretos por códigos, de modo a garantir a confidencialidade.

A integração com plataformas de pesquisa pode ser realizada por meio de *APIs* que permitem a transferência programada de dados. Para assegurar a segurança e a conformidade regulamentar, é imprescindível a implementação de protocolos adequados de forma a garantir que apenas utilizadores autorizados acedem às informações sensíveis. Estas medidas permitem não só proteger a privacidade dos doentes, mas também cumprir integralmente o RGPD, assegurando a gestão eficiente de dados de elevada qualidade e fortalecendo a base para pesquisas em saúde.

5. Conclusão

Este trabalho demonstrou que é possível retirar informação clínica através do cumprimento dos processos estabelecidos para a realização de *data mining* com qualidade.

Existe uma restrição a uma população muito específica de doentes que recorreram durante um período de tempo ao Grupo Lusíadas Saúde. Com este trabalho não é possível generalizar para a população portuguesa, mas destes resultados surge a possibilidade de realizar planos de ação dirigidos aos utilizadores específicos deste serviço de saúde. Serve de exemplo para realizar uma abordagem similar em outras bases de dados. Poder-se-ia assim comparar, com outros estudos, analisar incidências e taxas a um nível nacional. Hoje em dia facilitado pelo acesso a dados de qualidade e ao aumento exponencial da capacidade computacional é possível obter informação clínica relevante.

Determinaram-se patologias DALYs para a realização de planos de intervenção com locais de atuação delimitados a uma área específica, sem colocar em causa requisitos éticos ou de proteção de dados. Conclui-se que a utilização de dados de qualidade é fundamental para adquirirmos um maior conhecimento da população, possibilitando intervenções mais direcionadas e eficazes. No futuro a possibilidade de escalabilidade, de em tempo real os dados serem integrados numa plataforma digital, utilização de heatmaps por referência geográfica e a utilização de inteligência computacional serão uma mais valia para os profissionais e a comunidade que estes cuidam.

6. Referências

- [1] OCDE. Health in 21st century – Putting data to work for stronger health systems. OECD Health policy studies. 2021
- [2] Pearl R, Madvig P. Managing the most expensive patients. Harvard Business Review. 2020 Jan-Feb.
- [3] Direção-Geral da Saúde (DGS), Ministério da Saúde. Plano Nacional de Saúde 2021-2030. Saúde Sustentável: de tod@s para tod@s. 2022.
- [4] Coelho A, et al. Doenças não transmissíveis em Portugal: Desafios e oportunidades – Anais do IHMT
- [5] Ministério da Saúde, Direção-Geral da Saúde. Níveis de literacia em saúde - Portugal. 2021.
- [6] BRIEF project. Integrated brief interventions for noncommunicable disease risk factors in primary care: the manual. Copenhagen: WHO Regional Office for Europe; 2022.
- [7] Instituto Nacional de Estatística, I.P. Censos 2021 Resultados Definitivos – Portugal. 2022.
- [8] Graham F, Martin H, Lecouturier J, et al. Shared medical appointments in English primary care for long-term conditions: a qualitative study of the views and experiences of patients, primary care staff and other stakeholders. BMC Primary Care. 2022; 23:180.

- [9] Chen K., et al. Process mining and data mining applications in the domain of chronic diseases: A systematic review. *Artificial Intelligence In Medicine* 144. Agosto 2023
- [10] Hamdi H., et al. Enhancing Security and Privacy Preservation of Sensitive Information in e-Health Datasets Using FCA Approach. *ACM Computing Surveys*. Junho 2023.
- [11] Sarwar T., et al. The Secondary Use of Electronic Health Records for Data Mining: Data Characteristics and Challenges. *ACM Computing Surveys*, vol. 55, n. 2, Article 33. Janeiro 2022
- [12] Hunter D., Holmes C. Where Medical Statistics Meets Artificial Intelligence. *New England Journal Medicine* 2023;389:1211-9.
- [13] Nae T. et al. Covid severity prediction: Who cares about the data quality? *International Conference on Frontiers of Information Technology IEEE*. 2022
- [14] Rodrigues J., et al. Feature-Based Information Retrieval of Multimodal Biosignals with a Self-Similarity Matrix: Focus on Automatic Segmentation. *Biosensors*. 2022
- [15] Pandey A., Maneria S. Cloud Computing Methods Based on IoT for Better Patient Data Planning: A Research. *11th International Conference on System Modeling & Advancement in Research Trends*. 2022
- [16] Mirza N., et al. Leveraging the Internet of Things Aware Healthcare Monitoring System for Better Living. *Standards IEEE* 2022

- [17] Nagarajan S., et al. Intelligent Task Scheduling Approach for IoT Integrated Healthcare Cyber Physical Systems. IEEE Transactions on network science and engineering, vol. 10, nº. 5. Outubro 2023
- [18] Srinivasan K., et al. A Human-in-the-Loop Segmented Mixed-Effects Modeling Method for Analyzing Wearables Data. ACM Trans. Manage. Inf. Syst. 14, 2, Article 18. Janeiro 2023
- [19] Tao Y, et al. A Survey on Data-driven COVID-19 and Future Pandemic Management. ACM Computing Surveys, Vol. 55, No. 7, Article 141. Dezembro 2022.
- [20] Shahbazi N. et al. Representation Bias in Data: A Survey on Identification and Resolution Techniques. ACM Comput. Surv. 55, 13s, Article 293. Julho 2023
- [21] Fabris A., et al. Measuring Fairness Under Unawareness of Sensitive Attributes: A Quantification-Based Approach. Journal of Artificial Intelligence Research 76. Abril 2023
- [22] Kaur P., Nisa Z. Assessing Machine Learning Approaches for Imputing Missing Values in Cardiovascular Dataset. IEEE 7th International conference for Convergence in Technology (I2CT). Abril 07-09, 2022
- [23] Li Y., et al. Healthcare Data Quality Assessment for Cybersecurity Intelligence. IEEE – Transactions on industrial informatics, vol. 19, n. 1. Janeiro 2023
- [24] Bensina G., et al. Assessing the impact of kidney cancer-related premature mortality and productivity loss in Greece and Portugal. Expert review pharmaeconomics & outcomes research. 23:4 391-398. 2023

- 5] Brito J. et al. Atmospheric pollution and mortality in Portugal: Quantitative assessment of the environmental burden of disease using the AirQ+ model. *Science of the total environment*. 815. 2022
- [26] Nascimento B., et al. Carga de Doenças Cardiovasculares Atribuível aos Fatores de Risco nos Países de Língua Portuguesa: Dados do Estudo “Global Burden of Disease 2019”. *Arq Bras Cardiol*. 2022; 118(6):1028-1048
- [27] Lopes D., et al. Trajectories and determinants of ageing in Portugal: insights from EpiDoC, a nationwide population-based cohort. *BMC public health*. 2023; 23:1554
- [28] Rodrigues F., et al. Prospective study of loss of health-related quality adjusted life years in children and their families due to uncomplicated and hospitalised varicela. *Vaccine*. 41. 2022
- [29] Masters in Data Science. How to Deal with Missing Data. Disponível em: [How to Deal with Missing Data](<https://www.mastersindatascience.org/learning/how-to-deal-with-missing-data/>).
- [30] Dong Y, Peng CYJ. Principled missing data methods for researchers. *SpringerPlus*. 2013;2:222. Disponível em: [SpringerPlus](<http://www.springerplus.com/content/2/1/222>).
- [31] Sterne JAC, White IR, Carlin JB, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ*. 2009; 338:b2393. doi: <https://doi.org/10.1136/bmj.b2393>.

- [32] Roth GA, Johnson CO, Abajobir A, et al. Global, Regional, and National Burden of Cardiovascular Diseases for 10 Causes, 1990 to 2015. *Journal of the American College of Cardiology*. 2017;70(1):1-25. doi:10.1016/j.jacc.2017.04.052.
- [33] World Health Organization (WHO). *International Classification of Diseases, Ninth Revision (ICD-9)*. Geneva: World Health Organization; 1977. Disponível em: [WHO ICD-9](https://www.who.int/classifications/icd/en/).
- [34] Murray CJL, Lopez AD. *The Global Burden of Disease: A comprehensive assessment of mortality and disability from diseases, injuries, and risk factors in 1990 and projected to 2020*. Cambridge, MA: Harvard University Press; 1996.
- [35] Orso M, Cozzolino F, Amici S, De Giorgi M, Franchini D, Eusebi P, et al. (2020) Validity of cerebrovascular ICD-9-CM codes in healthcare administrative databases. The Umbria Data-Value Project. *PLoS ONE* 15(1): e0227653. <https://doi.org/10.1371/journal.pone.0227653>
- [36] Fiest et al. *BMC Psychiatry* 2014, 14:289 <http://www.biomedcentral.com/1471-244X/14/289>
Systematic review and assessment of validated case definitions for depression in administrative data
- [37] Cooke et al. *BMC Health Services Research* 2011, 11:37 <http://www.biomedcentral.com/1472-6963/11/37> The validity of using ICD-9 codes and pharmacy records to identify patients with chronic obstructive pulmonary disease

- [38] Manual of the international statistical classification of diseases, injuries and causes of death. Recomendações da nona conferencia de revisão 1975 - WHO 1977
- [39] Wallwork SB, Braithwaite FA, O’Keeffe M, Travers MJ, Summers SJ, Lange B, Hince DA, Costa LOP, Menezes Costa LC, Chiera B, Moseley GL. The clinical course of acute, subacute and persistent low back pain: a systematic review and meta-analysis. CMAJ. 2024 January 22;196:E29-46. doi: 10.1503/cmaj.230542.
- [40] Wu A, March L, Zheng X, Huang J, Wang X, Zhao J, Blyth FM, Smith E, Buchbinder R, Hoy D. Global low back pain prevalence and years lived with disability from 1990 to 2017: estimates from the Global Burden of Disease Study 2017. Transl Med. 2020;8(6):299.
- [41] Kislaya I, Neto M. Caracterização sociodemográfica da prevalência da dor lombar crónica autorreportada na população residente em Portugal através do Inquérito Nacional de Saúde 2014. Departamento de Epidemiologia, Instituto Nacional de saúde Doutor Ricardo Jorge, Lisboa, Portugal. Sociodemographic characterization of self-reported chronic low back pain prevalence in Portugal: results from the National Health Survey 2014.
- [42] Scholten M, Davidge J, Agvall B, Halling A. Comorbidities in heart failure patients that predict cardiovascular readmissions within 100 days—An observational study. PLoS ONE. 2024;19(1):e0296527.

- [43] Zhou D, Wang L, Ding S, Shen M, Qiu H. Phenotypic Disease Network Analysis to Identify Comorbidity Patterns in Hospitalized Patients with Ischemic Heart Disease Using Large-Scale Administrative Data. *Healthcare*. 2022;10:80.

7. Anexos

7.1 Pareceres das comissões de ética do Grupo Lusíadas Saúde

PARECER DA COMISSÃO DE ÉTICA PARA A SAÚDE PARA ESTUDOS OU ENSAIOS CLÍNICOS

1. Título do Estudo ou Ensaio Clínico

Data mining approaches for bio-demographic and morbidity analysis

2. Investigador Principal

Enfermeiro Pedro Nogueira

Parecer da Comissão de Ética para a Saúde do Hospital Lusíadas Porto

Analisados os documentos inerentes ao estudo "Data mining approaches for bio-demographic and morbidity analysis", tendo como Investigador Principal Enf. Pedro Nogueira, a Comissão de Ética para a Saúde é de parecer que nada há a opor à realização do projeto em causa, que parece ter grande interesse científico.

O Presidente da CES-HLP


Dr. José Neves

19/10/2023



Parecer

Data: 17 de outubro de 2023

Assunto: **Estudo LCK_16-2023 : "Data mining approaches for bio-demographic and morbidity analysis".**

Requerente: Dr. Pedro Nogueira

No seguimento do requerimento que foi feito a esta Comissão de Ética para a Saúde, exposição essa que foi objeto de reflexão por parte da mesma Comissão, vem por este meio a CES dar o seu parecer quanto ao estudo em epígrafe.

Após a devida análise e discussão em momento oportuno, determina a CES que não existem objeções éticas relativamente ao Estudo LCK_16-2023. **"Data mining approaches for bio-demographic and morbidity analysis".**

O estudo foi votado pelos seguintes membros da CES:

Dr. Abel Abejas,
Dr. José Pedro Vasconcelos
Dr. João Bismark – Relator principal
Enfª. Mafalda García
Dr. Pedro Correia

Resultado da votação:

Parecer favorável. Deliberação aprovada por unanimidade.

A Comissão de Ética para a Saúde

Assinado por: ÁBEL GARCIA ABEJAS
Num. de Identificação: 30718889
Data: 2023.10.18 08:05:14 -01'00

Hospital Lusíadas Lisboa
Rua Abílio Mendes, 1500-458 Lisboa
Tel.: 21 770 40 40 | www.lusíadas.pt

SEDE ADMINISTRATIVA

FACULDADE DE **MEDICINA**

FACULDADE DE **CIÊNCIAS**

