Arquitetura Analítica Federada para Predição de Fratura Osteoporótica: Uma Prova de Conceito

Mestrado em Ciências de Dados

Santiago Andrés Rodrigues Manica

Leiria, julho 2026

Arquitetura Analítica Federada para Predição de Fratura Osteoporótica: Uma Prova de Conceito

Mestrado em Ciências de Dados

Santiago Andrés Rodrigues Manica

Estágio realizado sob a orientação do Professor Doutor Carlos Grilo, do Professor Doutor Rolando Miragaia, da Professora Doutora Ana Rodrigues e do Professor Doutor Tiago Taveira Gomes.

Leiria, julho 2026

Originalidade e Direitos de Autor

O presente relatório de estágio é original, elaborado unicamente para este fim, tendo sido devidamente citados todos os autores cujos estudos e publicações contribuíram para o elaborar.

Dedicatória

À minha esposa Inês e aos meus sogros pela compreensão e apoio incondicional.

Aos meus orientadores por me guiarem e tornarem tudo mais fácil.

Ao Instituto Politécnico de Leiria pela aprendizagem, ferramentas e oportunidades.

Agradecimentos

Agradeço a todos os que tornaram este trabalho possível.

Aos meus orientadores, pelo acompanhamento, disponibilidade e rigor científico ao longo de todo o processo. Aos colegas e amigos que partilharam ideias, revisaram textos e tornaram este percurso mais rico. À minha família, pelo apoio constante e pela compreensão nos momentos de maior dedicação ao trabalho. A todos os que, direta ou indiretamente, contribuíram para que esta dissertação fosse possível, o meu sincero obrigado.

Resumo

A osteoporose representa um problema de saúde pública significativo, associado a fraturas de fragilidade que resultam em elevada morbilidade, mortalidade e custos económicos. A estratificação de risco tradicional, baseada em ferramentas como o FRAX, apresenta limitações na sua capacidade preditiva e generalização. Este trabalho propõe uma metodologia para a análise de dados de saúde em múltiplas instituições, focada na predição de fraturas osteoporóticas, que garante a reprodutibilidade e a segurança dos dados através da contentorização.

O objetivo principal foi desenvolver e validar uma arquitetura analítica contentorizada (Docker + OMOP-CDM v5.4), assente no paradigma *code-to-data* e preparada para implementação federada em múltiplas instituições, em conformidade com o RGPD. Secundariamente, pretendeu-se implementar, sobre dados sintéticos de alta fidelidade, um pipeline de modelação preditiva e identificação de perfis de risco de fratura osteoporótica em adultos ≥50 anos, como prova de conceito para futura aplicação em dados clínicos reais.

Foi desenvolvida uma infraestrutura analítica reprodutível e portátil, assente no paradigma *code-to-data*, que permite executar toda a análise de forma padronizada e segura em qualquer instituição de saúde, sem necessidade de transferência de dados sensíveis.

O *pipeline* realiza a extração e pré-processamento de dados (formato OMOP-CDM), a modelação preditiva supervisionada (Random Forest, Gradient Boosting) e a análise de *clustering* não supervisionado (K-Means) para identificação de perfis de risco. A metodologia "trazer o código aos dados" (*code-to-data*) garante que os dados dos pacientes nunca saem da infraestrutura hospitalar, em conformidade com o RGPD.

O resultado principal é um pacote de análise contentorizado, portátil e reprodutível, que permite a qualquer instituição com dados no formato OMOP-CDM executar um pipeline de análise de risco de fratura de forma padronizada. Os modelos de *machine learning* demonstraram capacidade para identificar fatores de risco complexos, enquanto o *clustering* permitiu a segmentação de pacientes em perfis de risco distintos. Validados sobre dados sintéticos, os modelos de *machine learning* demonstraram capacidade para identificar fatores de risco complexos e o *clustering* permitiu a segmentação em perfis de risco distintos. A arquitetura está preparada para implementação imediata com dados clínicos reais em ambiente federado.

A metodologia de contentorização demonstrou ser uma solução eficaz e segura para a análise distribuída de dados de saúde, superando barreiras de privacidade e promovendo a investigação colaborativa. Este trabalho estabelece uma base técnica robusta para a implementação de estudos multicêntricos e de aprendizagem federada em Portugal, com potencial para melhorar a estratificação de risco e a prevenção de fraturas osteoporóticas.

Palavras-chave: Osteoporose, Risco de Fratura, *Machine Learning*, SMOTE, Balanceamento de Classes, Análise Federada, Docker, Reprodutibilidade Científica

Abstract

Osteoporosis represents a significant public health problem, associated with fragility fractures that result in high morbidity, mortality, and economic burden. Traditional risk stratification, based on tools such as FRAX, has limitations in its predictive capacity and generalisability. This work proposes an innovative methodology for the analysis of health data across multiple institutions, focused on the prediction of osteoporotic fractures, ensuring reproducibility and data security through containerisation.

The primary objective was to develop and validate a containerised analytical architecture (Docker + OMOP-CDM v5.4), based on the code-to-data paradigm and prepared for federated deployment across multiple institutions, in compliance with GDPR. Secondarily, the aim was to implement, on high-fidelity synthetic data, a predictive modelling and fracture risk profiling pipeline for adults aged ≥50 years, as a proof of concept for future application with real clinical data.

A reproducible and portable analytical infrastructure was developed, based on the code-to-data paradigm, enabling standardised and secure execution of the full analysis pipeline at any health institution without the need to transfer sensitive patient data. The pipeline performs data extraction and preprocessing (assuming OMOP-CDM format), supervised predictive modelling (Random Forest, Gradient Boosting), and unsupervised clustering analysis (K-Means) for risk profile identification. The code-to-data methodology ensures that patient data never leave the hospital infrastructure, guaranteeing GDPR compliance.

The primary output is a containerised, portable, and reproducible analysis package that enables any institution holding data in OMOP-CDM format to run a standardised fracture risk analysis pipeline. Class imbalance (6.4% fracture events) was addressed using SMOTE applied within each cross-validation fold to prevent data leakage. Validated on synthetic data (Synthea), the machine learning models demonstrated the ability to identify complex risk factors, and clustering enabled the segmentation of patients into distinct risk profiles. The architecture is ready for immediate deployment with real clinical data in a federated setting.

The containerisation methodology proved to be an effective and secure solution for distributed health data analysis, overcoming privacy barriers and promoting collaborative research. This work establishes a robust technical foundation for the implementation of multicentre studies and federated learning in Portugal, with potential to improve risk stratification and the prevention of osteoporotic fractures.

Keywords: Osteoporosis, Fracture Risk, Machine Learning, SMOTE, Class Imbalance, Federated Learning, Docker, Scientific Reproducibility

Índice

Lista de Figuras

Lista de tabelas

Lista de siglas e acrónimos

ATC	Sistema de Classificação Anatómica Terapêutica Química (Anatomical Therapeutic Chemical)
AUC	Área sob a curva (Area Under the Curve)
BMI	Índice de Massa Corporal (Body Mass Index)
CI	Intervalo de Confiança (Confidence Interval)
DBSCAN	Density-Based Spatial Clustering of Applications with Noise)
DEXA	Absorciometria de Dupla Energia de Raios X (Dual-Energy X-ray Absorptiometry)
DGS	Direção-Geral da Saúde
DMO	Densidade Mineral Óssea
EHDEN	European Health Data and Evidence Network
EHR	Registo Eletrónico de Saúde (Electronic Health Record)
EXAM	Electronic Medical Record Chest X Ray Artificial Intelligence Model
FL	Aprendizagem Federada (Federated Learning)
FRAX	Ferramenta de Avaliação do Risco de Fratura (Fracture Risk Assessment Tool)
HDBSCAN	Hierarchical Density-Based Spatial Clustering of Applications with Noise
ICD-10	Classificação Internacional de Doenças, 10.ª Revisão
ICPC2	Classificação Internacional de Cuidados Primários versão 2
IMC	Índice de Massa Corporal
KM	Kaplan-Meier
MELLODDY	Machine Learning Ledger Orchestration for Drug Discovery
ML	Aprendizagem Automática (Machine Learning)
NHANES	National Health and Nutrition Examination Survey
OHDSI	Observational Health Data Sciences and Informatics
OMOP-CDM	Observational Medical Outcomes Partnership – Common Data Model
OR	Odds Ratio
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PROBAST	Prediction model Risk Of Bias ASsessment Tool
RCT	Ensaio Clínico Aleatorizado (Randomized Controlled Trial)
RF	Random Forest
RGPD	Regulamento Geral sobre a Proteção de Dados
RWD	Dados do Mundo Real (Real-World Data)
SQL	Structured Query Language
TI	Tecnologia de Informação
TRE	Ambiente de Investigação Fidedigno (Trusted Research Environment)

TTT	Treat-to-Target
ULS	Unidade Local de Saúde
VM	Máquina Virtual – Virtual Machine
XAI	Explainable AI
XGBoost	eXtreme Gradient Boosting

1. Introdução

A osteoporose é uma doença crónica e silenciosa, caracterizada pela perda progressiva de massa óssea e pela deterioração da microarquitetura do esqueleto. Como consequência, aumenta a fragilidade óssea e o risco de fraturas mesmo após traumatismos mínimos ou atividades quotidianas. Trata-se de uma condição prevalente no envelhecimento populacional e constitui um dos principais problemas de saúde pública, incluindo em Portugal [1]. As fraturas osteoporóticas, em particular as da anca e da coluna, estão associadas a dor crónica, perda de autonomia funcional, dependência a longo prazo e aumento da mortalidade. Para além do impacto clínico direto, estas fraturas acarretam uma carga económica significativa para os sistemas de saúde e para a sociedade, associada a hospitalizações prolongadas, intervenções cirúrgicas e programas de reabilitação [2].

Estima-se que a sua incidência continue a aumentar nas próximas décadas, acompanhando a transição demográfica do envelhecimento [3]. Ao longo das últimas décadas, foram desenvolvidos diversos modelos preditivos para estimar o risco de fratura osteoporótica. Entre os mais conhecidos encontram-se a Ferramenta de Avaliação do Risco de Fratura (FRAX, do inglês Fracture Risk Assessment Tool), o QFracture e o Garvan, que integram variáveis como idade, sexo, densidade mineral óssea e fatores de risco clínicos [4].

Estes modelos tiveram ampla difusão na prática clínica e representam um avanço importante na estratificação do risco. No entanto, apresentam limitações reconhecidas: simplificam variáveis clínicas complexas, não incorporam fatores como quedas recentes e são frequentemente validados em populações específicas, o que limita a sua generalização.

Nos últimos anos, os avanços em ciência de dados e em capacidade computacional abriram novas perspetivas. Algoritmos de *machine learning* (ML) demonstraram capacidade de identificar relações não lineares e interações complexas entre múltiplos fatores clínicos, laboratoriais, genéticos e de estilo de vida, alcançando desempenhos superiores aos modelos tradicionais [5]. Contudo, a literatura mostra que muitos destes modelos ainda carecem de validação externa robusta e enfrentam dificuldades de interpretabilidade, o que restringe a sua aplicação clínica direta [6]. Este movimento insere-se num paradigma mais amplo da medicina preditiva e personalizada, em que a utilização de grandes volumes de dados clínicos e biomédicos, aliados a técnicas de inteligência artificial, procura transformar a prática médica [7]. O recurso a dados de vida real provenientes de registos eletrónicos de saúde (RWD, do inglês Real-World Data) desempenha aqui um papel fundamental, pois permite criar modelos mais representativos e próximos da diversidade dos doentes seguidos em contexto clínico habitual.

O presente projeto procura responder a estas lacunas, propondo a construção e validação de modelos preditivos do risco de fratura osteoporótica em adultos com 50 ou mais anos, com recurso a dados sintéticos que simulam registos clínicos reais. Para tal, integra abordagens estatísticas tradicionais, algoritmos de ML e técnicas de *clustering* não supervisionado, com o objetivo de melhorar a estratificação de risco e apoiar a decisão clínica baseada em evidência.

1.1. Questão científica e objetivos

Este trabalho insere-se num projeto de investigação mais alargado que visa criar uma infraestrutura analítica para predição de fraturas osteoporóticas em dados clínicos reais de múltiplas Unidades Locais de Saúde (ULS) portuguesas. O princípio orientador dessa infraestrutura é o paradigma *code goes to data*: em vez de centralizar dados sensíveis num servidor de análise, o código analítico é distribuído e executado localmente em cada instituição, respeitando as restrições éticas e legais do RGPD. Na presente fase, o *pipeline* foi validado com dados sintéticos gerados pelo Synthea (MITRE Corporation [8]), mapeados para o OMOP-CDM v5.4 (*Observational Medical Outcomes Partnership Common Data Model*, um modelo de dados aberto desenvolvido pelo consórcio OHDSI para harmonizar registos clínicos de diferentes fontes [9]), servindo como representação sintética equivalente dos dados clínicos reais cuja disponibilização aguarda aprovação ética.

A questão central que orienta este trabalho é: é possível conceber e validar uma arquitetura analítica reprodutível e portátil, assente no paradigma *code-to-data* e preparada para execução federada, que permita predizer e estratificar o risco de fratura osteoporótica, combinando modelação preditiva supervisionada e a identificação de perfis clínicos por *clustering* não supervisionado, de forma segura e em conformidade com o RGPD, através de uma arquitetura de Aprendizagem Federada (*Federated Learning*), sem transferência de dados sensíveis entre instituições?

Objetivo principal: Desenvolver e validar uma arquitetura analítica reprodutível, assente em contentorização Docker e no paradigma *code-to-data*, preparada para implementação federada em múltiplas ULS em conformidade com o RGPD, com vista à modelação preditiva e identificação de perfis clínicos de risco de fratura osteoporótica em adultos com 50 ou mais anos. O objetivo científico final será concretizado quando os dados das ULS participantes forem injetados na mesma arquitetura, com eventuais ajustes limitados à camada de ingestão de dados, preservando a integridade dos módulos analíticos.

Objetivos secundários:

- Implementar e validar funcionalmente uma infraestrutura analítica contentorizada (Docker + OMOP-CDM v5.4) com dados sintéticos de alta fidelidade, demonstrando a sua prontidão para execução em ambiente hospitalar real;

- Estimar a prevalência de fraturas osteoporóticas e descrever fatores clínicos, demográficos e de estilo de vida associados;

- Construir modelos prognósticos com variáveis clínicas e comportamentais (idade, sexo, IMC, tabagismo, comorbilidades, medicação);
- Avaliar o desempenho preditivo de algoritmos tradicionais e de *machine learning*;
- Aplicar técnicas de *clustering* para identificar subgrupos com perfis de risco distintos.

1.2. Descrição Sumária dos Métodos

Este trabalho seguiu o desenho de um estudo observacional, retrospetivo e longitudinal (10 anos). Na presente fase, foi aplicado a dados sintéticos que simulam registos eletrónicos de saúde de adultos ≥50 anos, replicando a estrutura dos dados clínicos reais das ULS participantes.

O desfecho principal foi a ocorrência de fratura osteoporótica incidente (ICD-10: S72.0, M48.4–5, S52.5–6, M80.0). Foram ainda analisadas categorias de fratura (0, 1 ou ≥2 fraturas).

Foram incluídas variáveis demográficas, antropométricas, clínicas (ex. DEXA, FRAX), comorbilidades (Classificação Internacional de Doenças, 10.ª Revisão (ICD-10)/Classificação Internacional de Cuidados Primários versão 2 (ICPC2)) e terapêuticas Anatomical Therapeutic Chemical (ATC). Os dados foram analisados com técnicas de imputação, estatística descritiva, modelação preditiva (regressão, Random Forest, Gradient Boosting) e *clustering* (K-Means, Density-Based Spatial Clustering of Applications with Noise (DBSCAN), análise hierárquica). A comparação de risco entre clusters foi realizada com curvas de Kaplan-Meier.

1.3. Estrutura do trabalho

Os restantes capítulos do documento estão estruturados da seguinte forma: o Capítulo 2 apresenta a revisão da literatura; o Capítulo 3 descreve os métodos utilizados; o Capítulo 4 expõe os resultados; o Capítulo 5 discute os resultados e suas implicações; finalmente, o Capítulo 6 conclui com recomendações e perspetivas futuras.

2. Revisão da Literatura

O presente capítulo tem como objetivo realizar uma revisão crítica e abrangente da literatura científica relacionada com a predição do risco de fratura osteoporótica, com ênfase na comparação entre modelos tradicionais, técnicas de ML e estratégias baseadas em dados de vida real (RWD, do inglês Real World Data). Inicialmente, são descritos os modelos de predição mais utilizados na prática clínica, como o FRAX, QFracture e Garvan. É depois explorado o papel emergente das abordagens baseadas em ML, avaliando a sua capacidade de superar limitações dos métodos tradicionais e de integrar múltiplos fatores clínicos, laboratoriais e ambientais na estratificação do risco individual.

Este capítulo examina a utilização crescente de RWD na construção de modelos preditivos, destacando as suas vantagens e os desafios associados. São também apresentados estudos que aplicaram técnicas de *clustering* não supervisionado para identificar perfis clínicos latentes de risco de fratura como nova perspetiva na estratificação populacional.

2.1. Modelos preditivos tradicionais

Nas últimas décadas, foram desenvolvidos vários modelos preditivos tradicionais para estimar o risco de fratura osteoporótica a 10 anos, combinando idade, sexo, densidade mineral óssea (DMO) e fatores clínicos de risco. Destacam-se três modelos amplamente utilizados: FRAX, Qfracture e Garvan.

2.1.1. FRAX (Fracture Risk Assessment Tool)

O FRAX é uma ferramenta desenvolvida pela Universidade de Sheffield (Reino Unido) que recorre a fatores clínicos (e.g., fratura prévia, história familiar, uso de corticóides, tabagismo, álcool) para estimar a probabilidade de fratura major osteoporótica e de anca em 10 anos e pode ainda integrar, opcionalmente, a DMO do colo do fémur avaliada por densitometria [4]. Este índice foi também validado para a população de Portugal, FRAX-Pt [10].

Apesar do seu uso difundido, o FRAX tem limitações reconhecidas: trata variáveis clínicas como o uso de glucocorticoides de forma binária, exclui fatores como quedas ou DMO lombar, não distingue fraturas recentes de antigas, e foi principalmente validado em indivíduos não tratados [11], [12].

2.1.2. QFracture

Desenvolvido no Reino Unido com base em dados de cuidados primários, o QFracture estima o risco de fratura a 10 anos e incorpora mais comorbilidades (por ex., *diabetes tipo 1 e 2*, doenças reumáticas, uso de psicotrópicos, histórico de quedas) que o FRAX, mas não usa DMO [13]. No entanto, é complexo, pouco prático em ambientes com dados incompletos e tem validação geográfica limitada [14].

2.1.3. Garvan Fracture Risk Calculator

Criado a partir do estudo Dubbo (Austrália), o Garvan (Garvan Fracture Risk Calculator), calcula o risco de fratura a 5 e 10 anos incorporando idade, género, DMO e número de fraturas e quedas prévias. Pode ser mais informativo que o FRAX em doentes com historial de quedas frequentes. Ainda assim, tal como outros modelos tradicionais, tem limitações na precisão preditiva [15].

2.2. Modelos de *Machine Learning*

Os modelos de ML têm ganho destaque por captarem relações não lineares complexas entre múltiplas variáveis, incluindo dados não tradicionais (como genética, imagens e sensores) [7]. Técnicas como Random Forest, eXtreme Gradient Boosting (XGBoost) e redes neurais têm mostrado desempenho superior ao FRAX em vários estudos. Uma meta-análise de 53 estudos (15,2 milhões de pacientes) encontrou C-Index médio ~0,75, com sensibilidade 0,76–0,79 e especificidade 0,81–0,83 [5], superando os valores do FRAX (area under the curve (AUC) 0,65–0,70).

Modelos ML em registos suíços e no UK Biobank atingiram C-index de 0,83 para fratura da anca em dois anos, superando o FRAX (0,62) [16]. Em doentes oncológicos, um modelo XGBoost alcançou AUC 0,92, versus 0,86 do FRAX [17]. Um estudo recente combinou risco genético (de 1.103 polimorfismos) com variáveis do FRAX, obtendo C-index 0,80 e melhoria na reclassificação de risco (NRI ~22%) [18].

Contudo, em bases com dados limitados (ex.: administrativos), os modelos de ML não superaram métodos simples: um modelo logístico obteve AUC ~0,70, equivalente ao XGBoost, evidenciando que a qualidade dos dados é crítica [19].

Uma revisão sistemática recente (Sun et al., 2022) analisou 70 modelos e 138 validações externas. Apenas 31% dos modelos passaram por validação externa, poucos apresentaram excelente discriminação ou calibração adequada, e nenhum mostrou baixo risco de viés (segundo a ferramenta de avaliação de risco de viés *Prediction model Risk Of Bias ASsessment Tool* (PROBAST), destacando a necessidade de modelos mais robustos e validados [6].

A Tabela 1 resume as principais diferenças entre os modelos preditivos tradicionais, as abordagens supervisionadas de ML e as técnicas não supervisionadas (*clustering*), destacando os tipos de tarefa, a natureza das relações modeladas e a forma como os dados são utilizados.

Tabela 1 - Comparação entre modelos tradicionais (FRAX, QFracture, Garvan), ML supervisionado e ML não supervisionado

Critério	Modelos Tradicionais (FRAX, QFracture, Garvan)	ML Supervisionado (Random Forest, XGBoost, Redes Neuronais)	ML Não Supervisionado (K-Means, DBSCAN, Hierárquico)
Tipo de Tarefa	Predição direta de risco	Predição direta de risco	Descoberta de padrões e perfis clínicos
Relação Modelada	Relações lineares, fixas	Relações complexas e não lineares	Sem pré-definição; baseia-se em proximidade/similaridade
Número de Variáveis	Limitado (variáveis clínicas selecionadas)	Pode incluir dezenas a centenas de variáveis	Pode incluir múltiplas variáveis, contínuas e categóricas
Exemplos de Variáveis	Idade, sexo, T-índice, fratura prévia	Idade, sexo, densitometria, marcadores bioquímicos, genéticos, histórico de quedas	Idade, sexo, T-índice, comorbidades, estilos de vida
Objetivo Final	Estimar risco absoluto de fratura	Estimar risco absoluto com maior precisão	Identificar subgrupos de pacientes com características clínicas semelhantes
Interpretação Clínica	Alta (fórmulas conhecidas)	Moderada a baixa (necessidade de técnicas de Explainable Artificial Intelligence (XAI))	Média (interpretação baseada nas características dos clusters)
Necessidade de Dados	Dados clínicos essenciais, estruturados	Volume elevado de dados, estruturados e não estruturados	Grandes volumes de dados heterogéneos, com ou sem rótulos
Validação Externa	Frequentemente realizada	Em progresso; ainda limitada em muitas áreas	Raramente validado de forma independente
Exemplos de Performance	AUC ~0,65–0,70 (FRAX)	AUC ~0,75–0,92 dependendo da base de dados	Identificação de clusters clinicamente distintos, não avaliado por AUC
Aplicação Clínica Atual	Ampla (uso rotineiro)	Emergente (algumas provas de conceito)	Exploratório (principalmente em investigação)
Principais Desafios	Inflexibilidade, exclusão de variáveis relevantes	Interpretabilidade, integração nos fluxos de trabalho	Definição clínica e valida-

		clínico	ção de significado dos clusters
Exemplos Clínicos	FRAX-Pt, QFracture no Reino Unido	Modelos com UK Biobank, Medicare Data	Clusters de fragilidade vs. robustez em estudos dinamarqueses e canadianos

2.3. Utilização de Dados de Mundo Real

Os dados de mundo real (RWD, do inglês Real-World Data) tornaram-se um recurso essencial para a investigação em saúde, permitindo análises mais próximas da prática clínica habitual e complementando a evidência obtida em ensaios clínicos tradicionais.

2.3.1. Utilização de Dados de Mundo Real na Predição de Fraturas Osteoporóticas

A disponibilidade crescente de registos eletrónicos de saúde (EHR) tem impulsionado o desenvolvimento de modelos preditivos mais personalizados e representativos, possibilitando a inclusão de populações amplas e clinicamente complexas, frequentemente excluídas de ensaios clínicos [19], [20].

Têm sido utilizadas técnicas de ML para explorar estas bases com maior profundidade, identificando padrões complexos mesmo em contextos com dados incompletos. Alguns estudos demonstram que modelos desenvolvidos com dados clínicos ricos, como incluindo DMO, histórico de quedas, comorbilidades e exposição medicamentosa, podem atingir AUC superiores a 0,90, enquanto bases mais administrativas tendem a obter valores em torno de 0,65–0,70 [19], [21].

Para garantir análises válidas e replicáveis, é essencial padronizar a codificação de diagnósticos (por exemplo, segundo a Classificação Internacional de Doenças – ICD-10) e de medicamentos (pelo sistema Anatomical Therapeutic Chemical – ATC) [22],[23], [24].

A exposição farmacológica é frequentemente medida em Doses Diárias Definidas (DDD), permitindo estimar a dose cumulativa mesmo na ausência de dados sobre duração exata do tratamento. Um exemplo relevante foi demonstrado num estudo dinamarquês, onde uma dose cumulativa ≥1 g de prednisona se associou a risco quase três vezes superior de fratura do colo do fémur (Odds Ratio (OR) ~2,9) [25]. Métodos como o Weighted Cumulative Exposure (WCE) permitem ainda modelar o impacto temporal da exposição, atribuindo pesos diferentes às doses consoante a sua proximidade ao evento [26].

A qualidade e granularidade dos dados são determinantes para o desempenho dos modelos, influenciando tanto algoritmos supervisionados como a formação de *clusters* clinicamente relevantes. A heterogeneidade entre instituições, a ausência de variáveis-chave e os erros de codificação são desafios recorrentes [27], [28].

Além disso, a maioria dos modelos preditivos calcula o risco de forma estática, com base em valores basais. No entanto, o risco de fratura é dinâmico e evolui com o tempo. Torna-se, portanto, necessário desenvolver modelos adaptativos que incorporem novas informações à medida que o estado de saúde do paciente se altera. A aprendizagem contínua a partir de dados em fluxo provenientes dos EHR representa uma frente promissora, ainda pouco explorada.

Em resumo, os dados de mundo real oferecem uma base valiosa para melhorar a predição do risco de fratura osteoporótica, desde que acompanhados por padronização, rigor metodológico e estratégias que assegurem a sua qualidade e atualização constante [19], [27], [22], [25], [26],[28].

2.3.2. Modelação de risco e predição de desfechos com dados reais noutras áreas da saúde

O uso de dados do mundo real na modelação de risco e predição de desfechos tem-se expandido para várias áreas da saúde, refletindo a transversalidade destas abordagens. Em doenças neurodegenerativas, como a Doença de Parkinson, técnicas como o *clustering* têm identificado subgrupos clínicos com padrões distintos de sintomas e progressão. Alguns estudos demonstram que modelos de *machine learning* aplicados a dados reais conseguem prever indicadores como qualidade de vida, destacando os principais fatores associados [29]. Em cardiologia, oncologia e doenças metabólicas, como a diabetes, modelos preditivos e segmentação de pacientes com base em dados clínicos eletrónicos têm sido usados para estimar risco, personalizar terapêuticas e apoiar decisões clínicas [30], [31], [32], [33], [34].

2.4. Perfis de risco de fratura osteoporótica por *clustering* não supervisionado

Têm sido utilizadas técnicas de *clustering* para identificar perfis clínicos distintos de risco de fratura. Kruse et al. (2017), por exemplo, analisaram dados de mais de 10.000 mulheres e identificaram nove subgrupos com diferentes níveis de risco, com base em DMO, comorbilidades e idade. O estudo mostrou que o agrupamento hierárquico permite uma estratificação mais precisa do que o T-score isolado, revelando novos perfis de risco clínico [21].

Vincent et al. (2023) aplicaram análise de clusters a uma coorte canadiana, identificando dois perfis de indivíduos com fratura do punho: um mais idoso, frágil e com baixa DMO, e outro mais jovem, ativo e com DMO normal, cujas fraturas se deviam a trauma e não a fragilidade óssea [35]. Essa distinção tem relevância clínica, pois sugere que fraturas em participantes mais jovens e saudáveis podem não indicar o mesmo prognóstico nem necessitar das mesmas intervenções de prevenção secundária que fraturas em participantes frágeis [35].

Outros estudos de *clustering* focam-se nos fatores de risco subjacentes. Por exemplo, uma análise não supervisionada de dados do registo National Health and Nutrition Examination Survey (NHANES) com o algoritmo Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) identificou padrões inesperados nos fatores de risco tradicionais: idade avançada, uso diário de corticóides, história de fraturas prévias e antecedentes familiares emergiram como características principais nos clusters de maior densidade, ao passo que variáveis clássicas como sexo feminino, caucasianos ou menopausa precoce não se revelaram estatisticamente significativas em determinar os *clusters* de risco [36].

Estes resultados desafiam o estado da arte e sugerem que nem todos os fatores de risco conhecidos se manifestam de forma consistente nos dados de vida real, possivelmente devido à heterogeneidade populacional. Em conjunto, a literatura indica que as técnicas de *clustering* podem descobrir perfis clínicos latentes entre adultos ≥50 anos, diferenciando, por exemplo, um subgrupo de doentes com osteoporose e alta propensão a queda de outro subgrupo com ossos robustos, mas traumas de alta energia. Contudo, a utilidade clínica desses perfis ainda requer validação adequada [6].

2.5. Integração Clínica de *Machine Learning* e *Clustering*

A utilização de técnicas de ML e de *clustering* não supervisionado na previsão de fraturas osteoporóticas apresenta potencial para superar as limitações dos modelos tradicionais. Enquanto algoritmos como o FRAX consideram variáveis fixas e relações lineares, os modelos de ML são capazes de capturar interações complexas e não lineares entre múltiplos fatores clínicos, laboratoriais e de estilo de vida [37].

Além disso, os métodos de *clustering* permitem segmentar a população em perfis de risco mais heterogéneos, revelando subgrupos clinicamente relevantes que poderiam ser alvo de estratégias de prevenção diferenciadas. Esta abordagem personalizada poderá otimizar a identificação de indivíduos em risco, ultrapassando a visão simplificada baseada apenas em T-índice ou idade cronológica.

2.6. Limitações dos Estudos de *Clustering* na Predição de Fraturas Osteoporóticas

Apesar do potencial das técnicas de *clustering* não supervisionado para identificar perfis clínicos latentes de risco de fratura, existem várias limitações importantes a considerar. Em primeiro lugar, a heterogeneidade metodológica entre estudos, que aplicam métodos variados como *clustering* hierárquico, *two-step*, HDBSCAN ou *autoencoders*, dificulta a comparação direta dos perfis identificados. A qualidade e a granularidade dos dados clínicos constituem outro fator crítico, uma vez que bases de dados incompletas ou com variáveis limitadas podem gerar agrupamentos clinicamente pouco relevantes. Acresce que muitos *clusters* identificados são essencialmente descritivos, não tendo a sua

tradução em atitudes clínicas específicas, como a alteração de terapêutica com base no perfil de cluster, sido suficientemente validada. A generalização dos resultados é também limitada, dado que a maioria dos estudos foi realizada em populações específicas, como mulheres dinamarquesas ou canadianas, não sendo diretamente aplicável a outras populações sem revalidação. Paralelamente, a maior parte dos estudos reporta apenas validação interna dos clusters, sem recurso a conjuntos de dados independentes, o que compromete a robustez das conclusões. Por fim, permanece por demonstrar que o conhecimento do perfil de cluster melhora efetivamente desfechos clínicos, como a redução de fraturas, em comparação com estratégias de decisão tradicionais.

2.7. Considerações Éticas no Uso de Dados de Mundo Real (RWD)

A utilização de RWD para modelação preditiva, apesar das vantagens de representatividade populacional, levanta importantes questões éticas relacionadas com a proteção da privacidade e o consentimento informado. O Regulamento Geral sobre a Proteção de Dados (RGPD) estabelece que os dados de saúde são dados sensíveis e exige medidas robustas de anonimização ou pseudonimização antes do uso para investigação [38].

Adicionalmente, a integração de dados provenientes de diferentes fontes, como registos hospitalares, atenção primária e dispositivos *wearables*, reforça a necessidade de práticas éticas rigorosas para evitar riscos de reidentificação dos participantes.

2.8. Aprendizagem Federada

A análise de dados de saúde provenientes de múltiplas instituições, como as Unidades Locais de Saúde (ULS) em Portugal, enfrenta barreiras significativas relacionadas com a privacidade, segurança e governação de dados, especialmente no contexto do RGPD [39]. A partilha direta de dados de pacientes entre instituições é frequentemente impraticável. A Aprendizagem Federada (*Federated Learning*, FL) surge como um paradigma de aprendizagem automática distribuída que permite treinar modelos de forma colaborativa, sem que os dados brutos saiam das instalações onde se encontram [40], [41].

2.8.1. Conceitos e Arquitetura

A Aprendizagem Federada foi introduzida pela Google em 2016 como uma abordagem para treinar modelos em dados descentralizados, como os de dispositivos móveis [42]. O processo é tipicamente orquestrado por um servidor central e decorre de forma iterativa: o servidor define a arquitetura do modelo e inicializa os seus parâmetros, enviando-o de seguida para um conjunto de clientes participantes (e.g., hospitais, ULS). Cada cliente treina o modelo localmente com os seus dados, gerando uma atualização aos parâmetros, e envia apenas essa atualização para o servidor central, os dados

brutos nunca são partilhados. O servidor agrega então as atualizações recebidas para refinar o modelo global, sendo o algoritmo mais comum o Federated Averaging (FedAvg), que calcula uma média ponderada das atualizações dos clientes [42]. Este ciclo repete-se até que o modelo global convirja.

Esta abordagem "trazer o código aos dados" (*code goes to data*) minimiza a movimentação de dados sensíveis, reforçando a privacidade e a segurança, ao mesmo tempo que permite a criação de modelos mais robustos e generalizáveis, treinados em conjuntos de dados maiores e mais diversos [43].

Para além do FedAvg, outros algoritmos como o FedProx foram desenvolvidos para lidar com a heterogeneidade estatística (dados não-IID, non-independent and identically distributed) entre os diferentes clientes, um desafio comum em ambientes de saúde [44]. Para reforçar ainda mais a privacidade, técnicas como a Privacidade Diferencial (que adiciona ruído aos gradientes) e a Agregação Segura (que utiliza criptografia para que o servidor apenas consiga decifrar a soma das atualizações) podem ser integradas no processo [45], [46].

2.8.2. Aplicações na Saúde

A Aprendizagem Federada está a ganhar uma tração significativa na área da saúde, embora uma revisão sistemática recente tenha revelado que apenas uma pequena fração (5.2%) dos estudos publicados corresponde a aplicações em cenários de vida real [41]. As especialidades de radiologia e medicina interna são as mais comuns. As principais áreas de aplicação incluem:

Análise de Imagem Médica – A radiologia é a área com maior adoção, utilizando FL para tarefas como a segmentação de tumores cerebrais em imagens de ressonância magnética (MRI) [47], a classificação de lesões cutâneas [48] e a análise de retinografias [49], nomeadamente:

Análise de Registos Clínicos Eletrónicos (RCE): A FL permite criar modelos preditivos a partir de dados de RCE de múltiplos hospitais para prever a mortalidade de pacientes com COVID-19 [50], [51] ou o risco de hospitalização [52];

Genómica e Descoberta de Fármacos: A FL facilita estudos de associação genómica (GWAS) entre diferentes centros de investigação e permite que empresas farmacêuticas colaborem na descoberta de novos fármacos sem partilhar as suas bases de dados proprietárias, como demonstrado pelo projeto Machine Learning Ledger Orchestration for Drug Discovery (MELLODDY), que juntou 10 empresas farmacêuticas [53].

Um dos estudos de maior impacto em FL na saúde foi o estudo Electronic Medical Record Chest X-Ray AI Model (EXAM), que treinou um modelo para prever as necessidades futuras de oxigénio em pacientes com COVID-19. O estudo envolveu 20 instituições de cinco continentes e demonstrou que o modelo federado obteve uma melhoria de 16% na performance (AUC de 0.92) e 38% na generalização em comparação com modelos treinados localmente em cada instituição [50].

Este tipo de abordagem federada está a ser ativamente promovido na Europa através de iniciativas como a European Health Data & Evidence Network (EHDEN), que visa criar uma vasta rede de dados de saúde harmonizados para o modelo *Observational Medical Outcomes Partnership – Common Data Model* (OMOP-CDM), permitindo a realização de estudos observacionais em larga escala [54]. Em Portugal, investigadores como Taveira-Gomes têm tido um papel ativo no apoio à conversão de dados de parceiros de saúde portugueses para o formato OMOP-CDM, facilitando a sua integração em redes de investigação federadas [55]. Um exemplo é um estudo multinacional sobre cancro que analisou dados de 1.7 milhões de doentes de 11 bases de dados europeias, todas mapeadas para o OMOP-CDM, para estudar comorbilidades, uso de medicação e sobrevivência, demonstrando o poder da análise federada com dados de vida real [56]. Embora focados noutras áreas, estes trabalhos estabelecem um precedente metodológico e de infraestrutura diretamente aplicável a estudos multicêntricos sobre osteoporose em Portugal.

2.8.3. Aplicações em outras áreas

Fora do setor da saúde, a Aprendizagem Federada já é uma tecnologia estabelecida e utilizada em produção por grandes empresas de tecnologia com exemplificado na Tabela 2.

Tabela 2 - Aplicações atuais das metodologias federadas em diferentes áreas

Setor	**Aplicação**	**Exemplo**
Tecnologia	Previsão de texto em teclados móveis	Google Gboard [42]
Reconhecimento de voz e personalização	Reconhecimento de voz	Apple Siri, QuickType [55]
Finanças	Deteção de fraude e avaliação de crédito	WeBank (FATE Framework) [57]
Automóvel	Treino de modelos para veículos autónomos	NVIDIA, BMW [58]
Indústria	Controlo de qualidade e manutenção preditiva	Bosch [59]

2.8.4. Relevância da aprendizagem federada para o projeto

No contexto deste projeto, que visa analisar dados de múltiplas ULS em Portugal para identificar fatores de risco para fraturas osteoporóticas, a Aprendizagem Federada apresenta-se como uma solução metodológica robusta e alinhada com os requisitos do RGPD. A sua implementação permitiria treinar um modelo de risco de fratura mais preciso e generalizável, aproveitando a diversidade dos dados de diferentes unidades de saúde sem comprometer a privacidade dos pacientes. A abordagem metodológica detalhada no Capítulo 3, que utiliza contentores Docker, representa o primeiro passo

técnico para viabilizar uma futura análise federada, garantindo que o ambiente de análise é reprodutível e pode ser distribuído de forma segura por cada ULS participante.

2.9. Lacunas do conhecimento e justificação do projeto

Apesar do crescente volume de investigação dedicado à aplicação de algoritmos de *machine learning* e técnicas de *clustering* na predição de fraturas, a presente revisão identifica lacunas críticas que limitam a translação destes avanços para a prática clínica. A maioria dos modelos de alto desempenho foi desenvolvida em contextos controlados ou bases de dados específicas, carecendo de validação externa robusta em dados de vida real, e persiste o desafio da interpretabilidade clínica, sendo imperativo desenvolver soluções que garantam a transparência necessária para a tomada de decisão médica [37]. Por outro lado, os modelos convencionais e grande parte das abordagens de ML operam de forma estática, com base em avaliações pontuais, não capturando a natureza inerentemente dinâmica do risco de fratura ao longo do envelhecimento. A dispersão de dados por diferentes ULS e as restrições éticas e legais do RGPD constituem barreiras adicionais à criação de modelos robustos centralizados, embora a Aprendizagem Federada emerja como solução estratégica para contornar a fragmentação de dados, ainda com aplicação incipiente no domínio da saúde óssea em Portugal [41]. Acresce que, embora o *clustering* não supervisionado tenha demonstrado capacidade para identificar subgrupos de risco não captados por métricas tradicionais, a utilidade prática destes perfis na orientação de intervenções personalizadas não foi ainda validada em larga escala na população portuguesa. Este trabalho endereça estas lacunas através do desenvolvimento de uma arquitetura que utiliza dados de mundo real sob uma infraestrutura técnica assente em contentores Docker, garantindo a reprodutibilidade metodológica e preparando o terreno para implementações de aprendizagem federada multicêntrica.

3. Metodologia

A metodologia adotada no presente estudo organiza-se em diferentes componentes que abrangem o desenho do estudo, a caracterização da população e do período de observação, os procedimentos de recolha e processamento dos dados, bem como as estratégias estatísticas utilizadas para responder aos objetivos definidos.

3.1. Racional da Arquitetura: O Modelo de Duas Vias (Two-Track)

O desenvolvimento deste trabalho assenta na premissa de que a inovação em saúde digital não depende apenas da precisão dos modelos matemáticos, mas sobretudo da viabilidade da sua implementação em ambiente hospitalar real. Face às restrições éticas e legais (RGPD) que limitam a circulação de dados sensíveis, esta dissertação propõe uma solução tecnológica baseada no paradigma de "trazer o código aos dados" (*code goes to data*). A Figura 1 sumariza visualmente as diferenças entre o modelo clássico (os dados vão até o Cientista de Dados) e a análise proposta (*code goes to data*).

Para garantir a robustez e a prontidão desta solução, a metodologia foi estruturada em duas vias de execução paralelas, coordenadas de forma a tentar garantir a imutabilidade dos dados.

O desenvolvimento deste trabalho insere-se num contexto em que o acesso a dados clínicos reais está condicionado por restrições éticas, legais (RGPD) e operacionais, nomeadamente a ausência de aprovação ética final para acesso direto às bases de dados das ULS. Para não paralisar o desenvolvimento técnico durante este período de espera, a metodologia foi estruturada em duas vias de execução paralelas e independentes.

A Via 1 utiliza dados sintéticos gerados pelo simulador Synthea e tem como objetivo validar a integridade técnica do pipeline analítico; desde a ingestão de dados até à produção de modelos de risco e análise de sobrevivência. A Via 2 corresponde à aplicação futura do mesmo pipeline aos dados clínicos reais das ULS participantes, pendente de autorização ética.

O ponto central desta arquitetura é que os módulos de extração de conhecimento (NB01 a NB03) foram concebidos para máxima reutilização entre fontes de dados; contudo, a transição para dados reais da ULS implicará uma fase de revisão e ajuste, nomeadamente ao nível do módulo de carregamento (NB00) e da validação das variáveis extraídas. Esta separação não constitui uma garantia de portabilidade imediata: os dados reais poderão apresentar distribuições distintas, padrões de *missingness* diferentes e especificidades no mapeamento OMOP que exijam ajustes pontuais. O que a arquitetura assegura é que esses ajustes ficam confinados ao módulo de ingestão, preservando a integridade dos módulos analíticos.

Neste sentido, o principal resultado deste mestrado não é uma conclusão epidemiológica definitiva, mas a entrega de uma infraestrutura analítica validada com dados sintéticos de alta fidelidade e pronta para receber dados reais quando o acesso for autorizado.

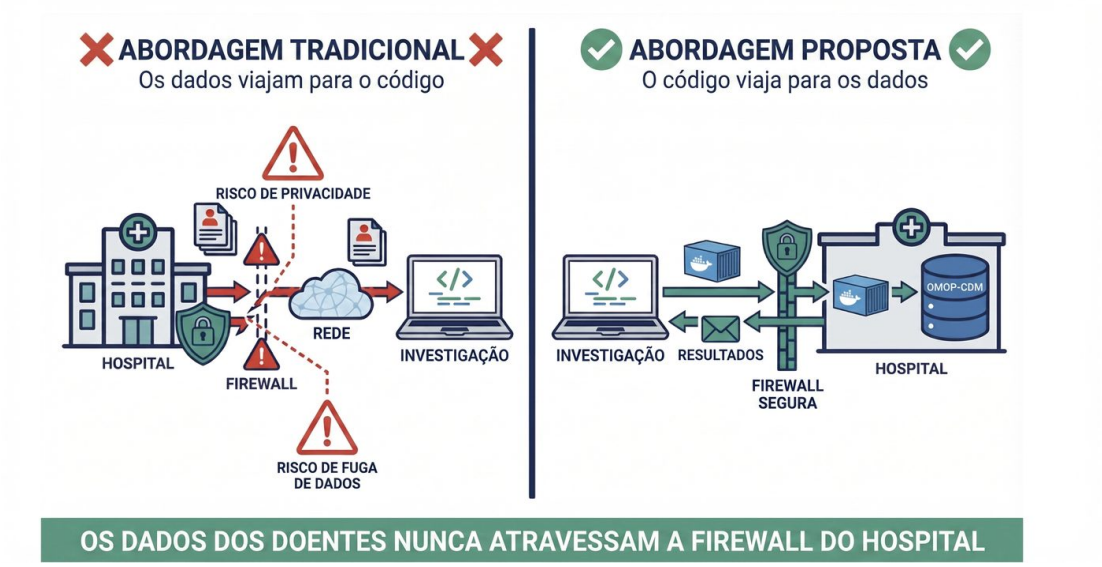

Figura 1 - Comparação entre a abordagem tradicional ("dados viajam para o código") e a abordagem adotada neste projeto ("código viaja para os dados"). Na abordagem federada, os dados dos pacientes nunca atravessam a firewall hospitalar

3.2. O Modelo de Dados Comum: OMOP-CDM

Para viabilizar a análise federada em dados de diferentes instituições, é crucial que os dados estejam num formato padronizado. Este projeto adota o Observational Medical Outcomes Partnership Common Data Model (OMOP-CDM) [9]. O OMOP-CDM é um padrão de dados aberto, desenvolvido pelo consórcio Observational Health Data Sciences and Informatics (OHDSI), que harmoniza a estrutura e o conteúdo de dados de saúde observacionais de diferentes fontes (e.g., registos eletrónicos de saúde, dados de faturação) numa representação comum [60].

3.3. Pipeline de Analítico

A análise está segmentada em três *Jupyter Notebooks*, cada um com uma responsabilidade específica, como se pode ver, esquematicamente, na Figura 2:

- 01_data_pull_and_feature_engineering.ipynb: Conecta-se à base de dados (PostgreSQL (Structured Query Language), extrai os dados em conformidade com o modelo OMOP-CDM, realiza a limpeza, o pré-processamento e a engenharia de *features*. Por simplicidade, este *notebook* é designado por NB01;

- 02_modeling_and_validation.ipynb: Utiliza os dados processados para treinar e validar os modelos de *machine learning* supervisionado (e.g., Random Forest, Gradient Boosting), avaliando a sua performance preditiva. Este *notebook* é designado por NB02;

- 03_clustering_and_survival.ipynb: Aplica técnicas de *clustering* não supervisionado para identificar perfis de risco e realiza análises de sobrevivência (Kaplan-Meier) para comparar os grupos identificados. Este *notebook* é designado por NB03.

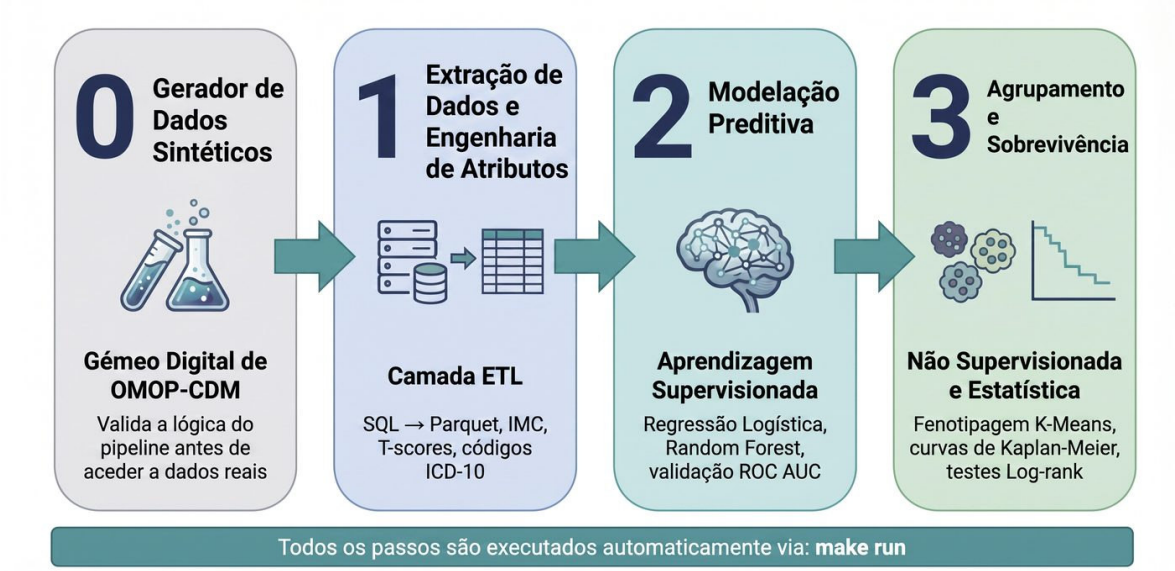

Figura 2 - Pipeline de análise de dados: desde a extração de dados (NB00) até ao *clustering* e análise de sobrevivência (NB03), orquestrado pelo comando make run

O pipeline inclui ainda um módulo preliminar, designado NB00 (00_data_preparation.ipynb), com uma função distinta dos restantes: a geração e ingestão de dados. Na Via 1 (validação), o NB00 é responsável pela simulação de dados sintéticos através do Synthea, populando a base de dados PostgreSQL no formato OMOP-CDM. Na Via 2 (produção com dados reais das ULS), este módulo é substituído pela ligação direta às fontes institucionais, não sendo necessária qualquer geração sintética. Os *notebooks* NB01 a NB03 permanecem inalterados em ambas as vias, a única diferença entre os dois cenários de execução reside no NB00.

3.4. Resumo do Fluxo de Dados

O fluxo integrado de ponta a ponta compreende duas fases distintas, executadas automaticamente via *make run*. Na primeira fase (NB00), o Synthea JAR gera 1.469 doentes sintéticos em formato CSV. Estes são filtrados por idade igual ou superior a 50 anos, resultando em 3.620 doentes, e mapeados para as 8 tabelas do OMOP-CDM v5.4. Nesta fase são ainda gerados estocasticamente os T-scores DXA e os valores laboratoriais, sendo os dados carregados num contentor Docker com PostgreSQL.

Na segunda fase (NB01), são extraídos 6 *Data Frames* em bruto (de acordo com a arquitetura

OMOP), totalizando 307.877 linhas. A coorte final mantém os 3.620 doentes. São então construídas 71 variáveis derivadas organizadas em 7 domínios clínicos, seguidas da definição dos *endpoints* de sobrevivência. Os valores em falta são imputados com imputação iterativa por regressão bayesiana.

O pipeline exporta dois ficheiros no formato Parquet: X_person.parquet, com 79 colunas correspondentes às variáveis clínicas e demográficas de cada doente, e surv.parquet, com 17 colunas contendo o tempo até ao evento e o indicador de fratura, utilizados nas análises de sobrevivência. Estes ficheiros alimentam as fases seguintes de modelação preditiva e agrupamento, nomeadamente os *notebooks* NB02 e NB03.

Esta arquitetura foi desenhada de forma a que eventuais ajustes na transição para dados clínicos reais se limitem à camada de ingestão de dados, preservando os módulos analíticos. Na prática, os dados reais poderão apresentar distribuições distintas, padrões de valores em falta diferentes e especificidades no mapeamento OMOP que exijam adaptações pontuais, o que a arquitetura procura minimizar, sem poder garantir ausência total de intervenção.

3.5. Desenho do estudo, população e período de estudo

A população-alvo incluiu adultos com idade igual ou superior a 50 anos. Os dados são representativos de um período de 10 anos (a partir de 2015). O *baseline* foi definido como a primeira consulta médica registada no sistema durante o período de análise, assegurando que as variáveis de exposição precedam a ocorrência de fraturas.

3.5.1. Fontes de Dados

Na presente fase, os dados utilizados foram gerados pelo simulador Synthea, que produz registos eletrónicos de saúde sintéticos de alta fidelidade, mapeados para o OMOP-CDM v5.4. O Synthea foi configurado para gerar uma coorte de adultos ≥50 anos, simulando as seguintes categorias de informação clínica:

- Dados demográficos (idade, género);
- Dados antropométricos e de estilo de vida (peso, altura, IMC, tabagismo, consumo de álcool);
- Resultados de densitometria óssea (T-score por DEXA) e estimativas do FRAX;
- Diagnósticos codificados (ICD-10 ou ICPC-2);
- Registos de prescrição medicamentosa (códigos ATC, dosagem e duração);
- Variáveis funcionais e de estilo de vida registadas na tabela *OBSERVATION* do OMOP-CDM: nível de atividade física (concept 4030093), risco de desnutrição (concept 4163261),

risco de queda (concept 4053525) e *score* de funcionalidade segundo as Atividades de Vida Diária – ADL (concept 4135496).

Foram ainda extraídas as seguintes variáveis clínicas adicionais: (i) *fratura prévia* – histórico pessoal de fratura documentado com data estritamente anterior ao *index date*; (ii) *Doença Renal Crónica (DRC),* identificada por códigos ICD-10 N18.1 a N18.6 e/ou por cálculo da taxa de filtração glomerular estimada (eGFR, equação CKD-EPI 2021), criando um *flag* binária e o estadio de DRC (0–5); (iii) *multimorbilidade* – variável binária que sinaliza a presença de duas ou mais comorbilidades crónicas, utilizada como variável de estratificação nas análises de subgrupo; (iv) *fármacos antiosteoporóticos* – além de uma *flag* geral de exposição a antiosteoporóticos, são criadas variáveis binárias por classe farmacológica (concept IDs OMOP): bifosfonato oral (1557272), bifosfonato IV (1524674), denosumab (40222444), romosozumab (35603812) e teriparatida/abaloparatida (1521987).

3.5.2 Definição dos Desfechos

Nos estudos observacionais e clínicos, um desfecho corresponde ao evento ou resultado de interesse a ser avaliado. O desfecho primário representa o principal resultado em análise, enquanto os desfechos secundários correspondem a resultados adicionais relevantes para a compreensão global do problema em estudo.

Desfecho primário:

Fratura incidente (variável binária), definida pelo primeiro registo de fratura osteoporótica durante o período de acompanhamento. Para uniformizar a identificação das fraturas, foi utilizada a ICD-10, que atribui códigos específicos a cada diagnóstico médico. Considerou-se como fratura incidente os novos registos com pelo menos um dos seguintes códigos:

- Fratura do colo do fémur: S72.0;
- Fratura vertebral: M48.4, M48.5;
- Fratura do punho: S52.5, S52.6;
- Fratura osteoporótica geral: M80.0;
- Fratura do úmero proximal: S42.2.

A primeira fratura documentada durante o *follow-up* foi considerada como evento incidente. Fraturas anteriores ao *index date* (data da primeira consulta registada) foram classificadas como antecedentes, garantindo a estrita precedência temporal das variáveis de exposição relativamente ao evento.

Desfecho secundário:

Categoria de fratura durante o *follow-up*. Enquanto o desfecho primário foca apenas na ocorrência

da primeira fratura, o desfecho secundário permite distinguir a gravidade do percurso clínico, classificando os doentes em:

- Sem qualquer fratura incidente;
- Uma fratura incidente;
- Duas ou mais fraturas (em momentos distintos).

3.6. Considerações Éticas

O estudo respeita integralmente o RGPD e os princípios éticos consagrados na Declaração de Helsínquia para a investigação médica envolvendo seres humanos. Ao adotar uma arquitetura federada, o protocolo incorpora o princípio de Privacidade desde a Conceção (*Privacy by Design*).

Neste modelo, os dados a nível do paciente permanecem estritamente sob a custódia e dentro das *firewalls* das respetivas instituições de saúde, garantindo a soberania institucional dos dados. Em vez da transferência de registos clínicos sensíveis, apenas resultados estatísticos agregados e totalmente anonimizados (ou atualizações de parâmetros matemáticos do modelo) são partilhados com o centro coordenador.

Esta abordagem assegura uma rigorosa minimização de dados, extraindo apenas o conhecimento estritamente necessário para responder à questão de investigação, anulando o risco de reidentificação durante o trânsito de informações. Adicionalmente, esta rede multicêntrica promove a equidade e a representatividade dos resultados científicos, mitigando vieses algorítmicos ao incluir populações diversificadas, sem nunca comprometer o direito à confidencialidade individual. O protocolo completo foi submetido à comissão de ética competente, estando a aprovação formal em curso.

3.7. Recolha de dados

Na fase atual do projeto, a extração de dados reais das ULS ainda não foi realizada (Track 2). O pipeline analítico foi desenvolvido e validado funcionalmente com dados sintéticos gerados pelo simulador Synthea [61], estruturados em formato OMOP-CDM v5.4 através do pacote oficial ETL-Synthea mantido pelo consórcio OHDSI [OHDSI ETL-Synthea; AWS Open Data]. O Synthea foi independentemente validado com quatro medidas de qualidade clínica, demonstrando fiabilidade para dados demográficos e percursos de cuidados [62]. Esta abordagem, designada Track 1, permite testar e depurar a cadeia analítica (NB01–NB03), validando a sua integridade técnica. Importa sublinhar, contudo, que cada conjunto de dados real pode exigir adaptações específicas ao pipeline, nomeadamente ao nível da extração e harmonização de variáveis (NB00). A ativação do módulo de dados reais pressupõe uma fase de revisão e ajuste. Os resultados apresentados nesta dissertação

referem-se, portanto, a esta fase de validação técnica e não a conclusões clínicas sobre a população portuguesa.

3.8. Pipeline Analítico: Organização e Fluxo de Dados

O tratamento e análise de dados foram organizados em quatro módulos sequenciais (NB00–NB03), implementados como Jupyter Notebooks numa infra-estrutura contentorizada. Estes módulos foram concebidos para poderem ser reutilizados; contudo, a transição para dados reais da ULS implicará uma fase de revisão e eventual adaptação. A Tabela 3 resume a função e o fluxo de dados de cada módulo.

Tabela 3. Organização do pipeline analítico e fluxo de dados entre módulos

Módulo	Função principal	Entradas	Saídas
NB00 — Carregamento e ETL	Mapeamento dos dados brutos para 8 tabelas OMOP-CDM v5.4; geração de variáveis laboratoriais estocásticas (T-score, vitamina D, DMO)	Synthea CSV (Track 1) ou dados ULS OMOP-CDM (Track 2)	PostgreSQL: 8 tabelas OMOP
NB01 — Extracção e Engenharia de Variáveis	Extracção das tabelas OMOP; construção de 71 variáveis derivadas em 7 domínios clínicos; endpoints de sobrevivência; imputação de missings	PostgreSQL (8 tabelas OMOP)	X_persons.parquet (79 colunas); surv.parquet (17 colunas)
NB02 — Modelação Preditiva	Treino e validação de modelos supervisionados (Log. Reg., RF, GBM, Cox); selecção de variáveis (RFECV); interpretabilidade (SHAP)	X_persons.parquet; surv.parquet	Métricas AUC-ROC; coeficientes HR; valores SHAP
NB03 — Clustering e Sobrevivência	Identificação de perfis clínicos por K-Means + PCA; validação por análise de sobrevivência (Kaplan-Meier, Cox)	X_persons.parquet; surv.parquet	Grupos clínicos; curvas KM por grupo; razões de risco

3.9. Pré-processamento de dados

O módulo NB01 inicia-se com a extração direta das oito tabelas OMOP-CDM carregadas pelo NB00 (PERSON, MEASUREMENT, CONDITION_OCCURRENCE, DRUG_EXPOSURE, OBSERVATION, VISIT_OCCURRENCE, OBSERVATION_PERIOD e DEATH).

A partir destas tabelas, são construídas 71 variáveis derivadas organizadas em sete domínios:

- (i) demográficos (idade na data-índice, sexo);
- (ii) antropométricos e de saúde óssea (IMC, T-score médio, DMO femoral e lombar);
- (iii) comorbilidades, identificadas por códigos ICD-10 nas tabelas CONDITION_OCCURRENCE;
- (iv) medicação antiosteoporótica, derivada da tabela DRUG_EXPOSURE com mapeamento por códigos ATC para classes como bifosfonatos, denosumab e teriparatida;
- (v) estilo de vida e função, extraídas da tabela OBSERVATION;
- (vi) variáveis calculadas, nomeadamente a TFGe pela equação CKD-EPI 2021, a categoria T-score (normal/osteopénia/osteoporose) e a *flag* de multimorbilidade;
- (vii) *endpoints* de sobrevivência (tempo ao evento e alvos multi-horizonte a 1, 3, 5 e 10 anos).

O pré-processamento inclui ainda uma estratégia formalizada de imputação, avaliando três métodos por validação cruzada de cinco grupos, e a normalização de variáveis contínuas para garantir a estabilidade dos algoritmos de aprendizagem automática.

3.9.1. Imputação de variáveis ausentes

Caso existam dados ausentes (*missing data*), relevante com dados reais, a omputação de valores ausentes é feita por seleção formal entre três métodos avaliados por validação cruzada de 5 grupos (5-fold CV RMSE): mediana estratificada por sexo com referências NHANES (método A), *imputação por k vizinhos mais próximos* (*k*=5, método B) e *IterativeImputer* (BayesianRidge, método C). O método com menor RMSE médio nos dados observados é automaticamente selecionado e aplicado ao conjunto completo de dados.

3.9.2. Normalização de variáveis contínuas

As variáveis contínuas incluídas na matriz analítica foram normalizadas por normalização z-score, resultando em média zero e desvio-padrão unitário. Este passo foi aplicado após a imputação e antes da modelação, de forma a garantir comparabilidade entre variáveis com escalas distintas (e.g., idade em anos vs. T-score) e a estabilizar a convergência dos algoritmos sensíveis à escala, como a Regressão Logística. As variáveis categóricas com múltiplas categorias (tabagismo, consumo de álcool,

categoria de IMC, categoria de T-score, estado de vitamina D e categoria de eGFR) foram codificadas por *one-hot encoding*, gerando colunas binárias para cada categoria. As variáveis binárias já codificadas como 0/1 foram tratadas como numéricas.

3.10. Seleção de Variáveis Preditoras (RFECV)

A seleção de variáveis foi realizada por Eliminação Recursiva com Validação Cruzada (RFECV), utilizando AUC-ROC como critério de otimização em validação cruzada estratificada de 5 grupos. O processo foi aplicado às variáveis numéricas da matriz analítica, excluindo variáveis binárias e categóricas já codificadas. O horizonte temporal alvo foi selecionado dinamicamente, exigindo um mínimo de 10 eventos nos dados de teste.

Complementarmente, foram aplicadas três técnicas de quantificação da importância das variáveis após o treino. Os valores SHAP (SHapley Additive exPlanations), baseados na teoria dos jogos cooperativos de Shapley, atribuem a cada variável uma contribuição marginal justa para cada predição individual; foram calculados para o melhor modelo global e separadamente por sexo, permitindo identificar se os preditores dominantes diferem entre homens e mulheres, informação clinicamente relevante numa doença com fisiopatologia marcadamente sexo-dependente. A importância por permutação (Permutation Importance), agnóstica ao modelo, quantifica o decréscimo no AUC-ROC quando os valores de cada variável são aleatoriamente permutados, servindo como validação independente dos valores SHAP. Por fim, os Partial Dependence Plots (PDP) mostram como a probabilidade predita de fratura varia em função dos valores das variáveis mais importantes, mantendo as restantes constantes, permitindo detetar não-linearidades e limiares clínicos relevantes.

3.11. Análise descritiva

Ainda no módulo NB01, após a construção do conjunto analítico, foi realizada uma análise descritiva completa da coorte. Esta etapa cumpriu dois objetivos: verificar a plausibilidade clínica dos dados sintéticos Synthea e fornecer o contexto epidemiológico necessário para interpretar os resultados dos modelos preditivos e do *clustering* (NB02 e NB03). A distribuição das variáveis (médias, desvios-padrão e frequências) foi avaliada globalmente e estratificada por sexo, por ser um modificador de risco reconhecido na osteoporose: as mulheres apresentam maior prevalência de osteoporose e de fratura vertebral, enquanto os homens registam pior prognóstico após fratura da anca. Os testes de comparação foram selecionados em função da escala de medição (t-test para variáveis contínuas e qui-quadrado para variáveis binárias/categóricas) permitindo identificar diferenças de base com relevância clínica entre grupos.

A análise descritiva compreendeu estatísticas descritivas gerais e testes de comparação entre grupos,

selecionados em função da escala de medição, t-test para variáveis contínuas e qui-quadrado para variáveis binárias e categóricas. A análise foi ainda estratificada por sexo (Tabela 1b), apresentando todas as variáveis demográficas, antropométricas, clínicas e farmacológicas separadamente para homens e mulheres, permitindo identificar diferenças de base com relevância clínica entre os dois grupos.

3.12.Modelação preditiva supervisionada

O módulo NB02 recebe os ficheiros X_persons.parquet e surv.parquet produzidos pelo NB01 e implementa a modelação preditiva supervisionada do risco de fratura. A estratégia assentou na comparação de cinco algoritmos com características complementares: a regressão logística como linha de base interpretável; o Random Forest e o Gradient Boosting para capturar relações não lineares e interações entre variáveis; e o XGBoost pela sua regularização nativa (L1 + L2) e robustez documentada em dados clínicos tabulares; e uma rede de perceptrões multicamada (MLP), incluído como baseline de rede neuronal rasa para referência comparativa.

O modelo de Cox de riscos proporcionais foi incluído para integrar a dimensão temporal do seguimento longitudinal. Antes do treino, a seleção de variáveis foi formalizada por RFECV com validação cruzada de cinco grupos. Todos os modelos foram avaliados com AUC-ROC como métrica primária, complementada pelo Brier score para avaliar a calibração. A abordagem multi-horizonte (1, 3, 5 e 10 anos) permite fornecer estimativas de risco acionáveis em diferentes janelas temporais.

A interpretabilidade dos modelos foi garantida pelo cálculo de valores SHAP, tanto globalmente como estratificado por sexo.

Divisão da Amostra (Treino/Teste)

A amostra foi dividida aleatoriamente com estratificação pelo evento de fratura (80% treino/validação, 20% teste), garantindo que a proporção de eventos se mantém em ambas as partições. A divisão aleatória estratificada é metodologicamente adequada neste contexto porque cada doente constitui uma observação independente: o histórico clínico de um doente não influencia o de outro, pelo que não existe dependência temporal entre observações que justifique uma divisão cronológica. A aleatoriedade maximiza o número de eventos no conjunto de teste, aumentando a fiabilidade das métricas de avaliação final.

AUC-CV: média das cinco áreas sob a curva ROC obtidas nas iterações do 5-fold cross-validation. Em cada iteração, quatro *folds* são usados para treino e um para validação interna; o processo repete-se cinco vezes de forma que todos os dados de treino sirvam uma vez como validação. É a métrica principal de seleção e comparação de modelos.

AUC-Teste: área sob a curva ROC calculada uma única vez sobre o conjunto de teste retido (≈20%

da amostra), que nunca participou em nenhuma etapa de treino nem de validação cruzada. Representa a estimativa de generalização do modelo final a dados completamente novos.

AUC-PR: área sob a curva Precision-Recall calculada no conjunto de teste. Em dados desequilibrados, a AUC-PR é mais informativa do que a AUC-ROC para avaliar a capacidade de identificar eventos raros. A linha de base corresponde à prevalência da fratura na amostra de teste (≈6,0%).

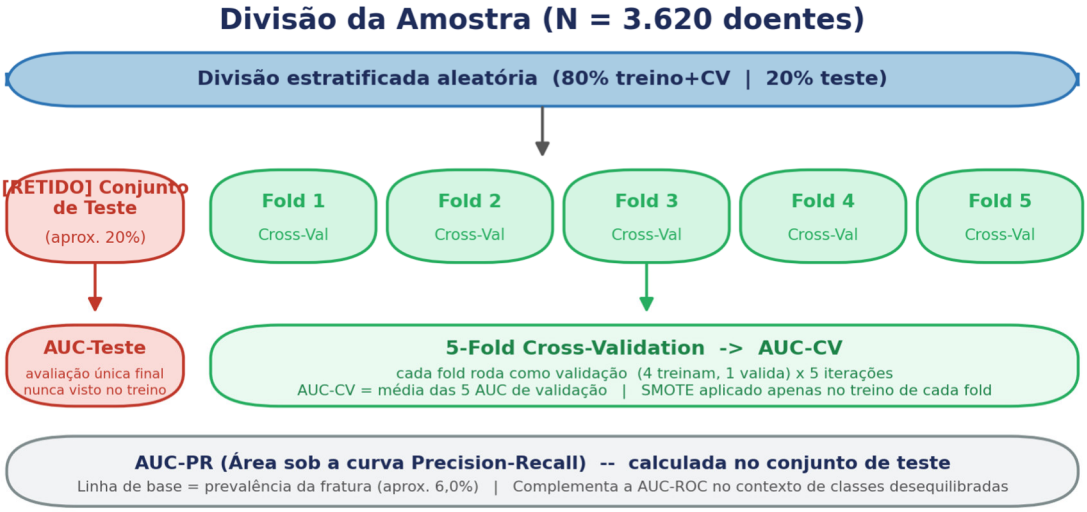

Figura 3. Esquema de divisão da amostra: o conjunto de teste (≈20%) é retido e bloqueado durante todo o processo de treino; os restantes dados são divididos em 5 folds para validação cruzada. A AUC-CV corresponde à média das 5 AUC de validação; a AUC-Teste é calculada uma única vez no conjunto retido; a AUC-PR quantifica o desempenho na classe minoritária.

3.13. Tratamento do Desequilíbrio de Classes (SMOTE)

O conjunto de dados apresenta um marcado desequilíbrio de classes: 219 fraturas (6,0%) versus 3.401 não-fraturas (94,0%). Para mitigar o enviesamento dos modelos para a classe maioritária, foi aplicada a técnica SMOTE (*Synthetic Minority Oversampling TEchnique*), uma forma de sobreamostragem que gera exemplos sintéticos da classe minoritária por interpolação linear entre vizinhos mais próximos ($k = 5$). A estratégia de amostragem adoptada foi proporcional 1:1 (sampling_strategy='auto'), igualando o número de exemplos das duas classes no conjunto de treino.

Criticamente, o SMOTE foi aplicado exclusivamente ao conjunto de treino de cada *split* de validação cruzada, através de um ImbPipeline (imblearn.pipeline), prevenindo a contaminação do conjunto de validação com dados sintéticos e garantindo a validade das estimativas de desempenho.

3.14. *Clustering* não supervisionado

O módulo NB03 aplica K-Means com redução dimensional prévia por PCA às variáveis selecionadas pelo RFECV. O número ótimo de *clusters* K é determinado pelo método do cotovelo e coeficiente de silhueta; *clusters* heterogéneos (silhueta < 0 ou dimensão < 5% da coorte) são excluídos antes da análise de sobrevivência. A validade clínica dos agrupamentos é aferida com curvas de Kaplan-Meier e modelo de Cox por *cluster*. O NB03 repete a análise separadamente para homens e mulheres (*clustering* intra-sexo), com seleção automática do K ótimo por sexo; os rótulos não são comparáveis entre grupos biológicos.

4. Resultados

Os resultados apresentados neste capítulo referem-se à execução completa do pipeline analítico (NB00–NB03) sobre dados sintéticos gerados pelo simulador Synthea (N=3.620 doentes; Track 1), estruturados em formato OMOP-CDM v5.4. O objetivo desta fase é validar funcionalmente a cadeia analítica, desde a extração de dados até à modelação preditiva e ao *clustering*, garantindo que o pipeline foi validado tecnicamente com dados sintéticos. A transição para dados clínicos reais das ULS (Track 2) exigirá uma fase de revisão e adaptação do pipeline, após aprovação ética.

Os resultados estão organizados em quatro blocos: (i) caracterização descritiva da coorte e dos desfechos (4.1); (ii) desempenho dos modelos preditivos supervisionados e importância das variáveis (4.2–4.4); (iii) segmentação não supervisionada por clustering K-Means, incluindo análise intra-sexo (4.5–4.6); e (iv) análise de sobrevivência por grupo e modelo de riscos proporcionais de Cox (4.7). Importa sublinhar que os valores numéricos reportados refletem padrões do gerador sintético e não constituem conclusões clínicas sobre a população portuguesa.

4.1. Características da Coorte

A coorte final compreendia 3.620 doentes com idade ≥50 anos. A distribuição por sexo estava equilibrada: 1.802 do sexo masculino (49,8%) e 1.818 do sexo feminino (50,2%). Um total de 219 doentes (6,0%) experienciaram pelo menos uma fratura osteoporótica incidente durante o *follow-up*. O *follow-up* mediano foi de 3.454 dias (9,5 anos).

Tabela 4 - Características basais da coorte

Característica	Valor	Notas
Total de doentes	3620	Idade ≥50 anos
Sexo (M / F)	1.802 (49,8%) / 1.818 (50,2%)	Equilibrado
Fraturas incidentes	219 (6,0%)	Endpoint primário
Follow-up mediano	3.454 dias (9,5 anos)	IQR: 1.257–4.017
Hipertensão	1.701 (47,0%)	Comorbilidade mais prevalente
Diabetes tipo 2	624 (17,2%)	
DRC (N18.3)	720 (19,9%)	
Neoplasia sólida / cancro	153 (4,2%)	
Multimorbilidade (≥2)	970 (26,8%)	

4.2. Pré-processamento e Engenharia de Variáveis

A construção da matriz de variáveis foi realizada no módulo NB01 a partir das oito tabelas OMOP-CDM. A engenharia de variáveis produziu uma matriz de 79 covariáveis por doente, abrangendo dados demográficos, comorbilidades codificadas em ICD-10/SNOMED, medições laboratoriais agregadas numa janela de 12 meses anterior à data de índice, medicações e variáveis derivadas (e.g., multimorbilidade, estadio de DRC, categoria de IMC).

A imputação de valores em falta nas variáveis laboratoriais e antropométricas foi realizada por validação cruzada de cinco folds, comparando três métodos: imputação iterativa por regressão Bayesiana (Iterative BayesRidge, RMSE=3,23), KNN com k=5 (RMSE=3,15) e mediana estratificada (RMSE=5,20). O método iterativo Bayesiano foi selecionado, resultando em zero valores em falta após imputação. A variável malnutrition_score foi excluída por apresentar 64,1% de valores em falta. A prevalência de prior_fracture foi de 0,5%, abaixo dos 5–10% esperados em dados reais, limitação conhecida do Synthea.

A seleção final de variáveis por RFECV reduziu de 64 variáveis numéricas para 7 ótimas (AUC CV=0,691): idade, IMC, altura, peso, creatinina, eGFR e estadio de DRC.

4.3. Análise Exploratória dos Dados

A distribuição etária da coorte situa-se entre os 50 e os 77 anos (média 60,7 ± 9,3 anos), com distribuição equilibrada por sexo (49,8% masculino). As comorbilidades mais prevalentes foram hipertensão (47,0%), diabetes tipo 2 (17,2%) e doença renal crónica (19,9%). O desequilíbrio de classes, 219 fraturas (6,0%) versus 3.401 sem fratura (94,0%), foi documentado e considerado na estratégia de modelação.

As características basais estratificadas por ocorrência de fratura estão resumidas na Tabela 1, que compara os grupos fratura (N=219) e sem fratura (N=3.401). A análise de correlação entre as 11 variáveis selecionadas pelo RFECV não identificou colinearidade severa, validando a estabilidade dos coeficientes dos modelos subsequentes.

4.4. Desempenho dos Modelos Preditivos

No Notebook 2 (NB02) foi avaliado o desempenho dos modelos preditivos.

O alvo de predição de fratura incidente (*event*) foi utilizado, com uma taxa de eventos de 6,0% (219 eventos em 3.620 doentes; divisão estratificada 80/20: treino n=2.896 com 175 eventos, teste n=724 com 44 eventos). O RFECV selecionou 7 variáveis (age_at_index, bmi_mean_12m, height_m, weight_kg, creatinine, egfr, ckd_stage). O desempenho dos modelos está resumido na Tabela 5.

Tabela 5. Comparação de modelos para predição de fratura a 10 anos.

Modelo	AUC-CV	AUC-Teste	AUC-PR	Brier	Melhor?
Regressão Logística	0,697	0,739	0,139	0,055	
Random Forest	0,732	0,771	0,128	0,055	
Gradient Boosting	0,712	0,779	0,130	0,055	
XGBoost	0,721	0,789	0,137	0,055	✓
MLP	0,720	0,721	0,107	0,057	

Os testes par a par de DeLong não encontraram diferenças estatisticamente significativas entre qualquer par de modelos (todos p>0,05). Com o *split* estratificado, os valores de AUC-Teste variam entre 0,721 (MLP) e 0,789 (XGBoost), refletindo a adequada separação de classes com 44 eventos no conjunto de teste. O XGBoost obteve o melhor desempenho sem SMOTE (AUC-Teste=0,789; AUC-PR=0,137), seguido do Gradient Boosting (AUC-Teste=0,779) e do Random Forest (AUC-Teste=0,771). O desempenho visual dos modelos está representado na Figura 4. A matriz de confusão do XGBoost (limiar de Youden ótimo=0,069) revela o impacto do desequilíbrio de classes: dos 44 casos de fratura no conjunto de teste, o modelo identifica 39 (sensibilidade=88,6%), com 248 falsos positivos (precisão=13,6%). O gráfico de calibração confirma discriminação adequada para dados sintéticos.

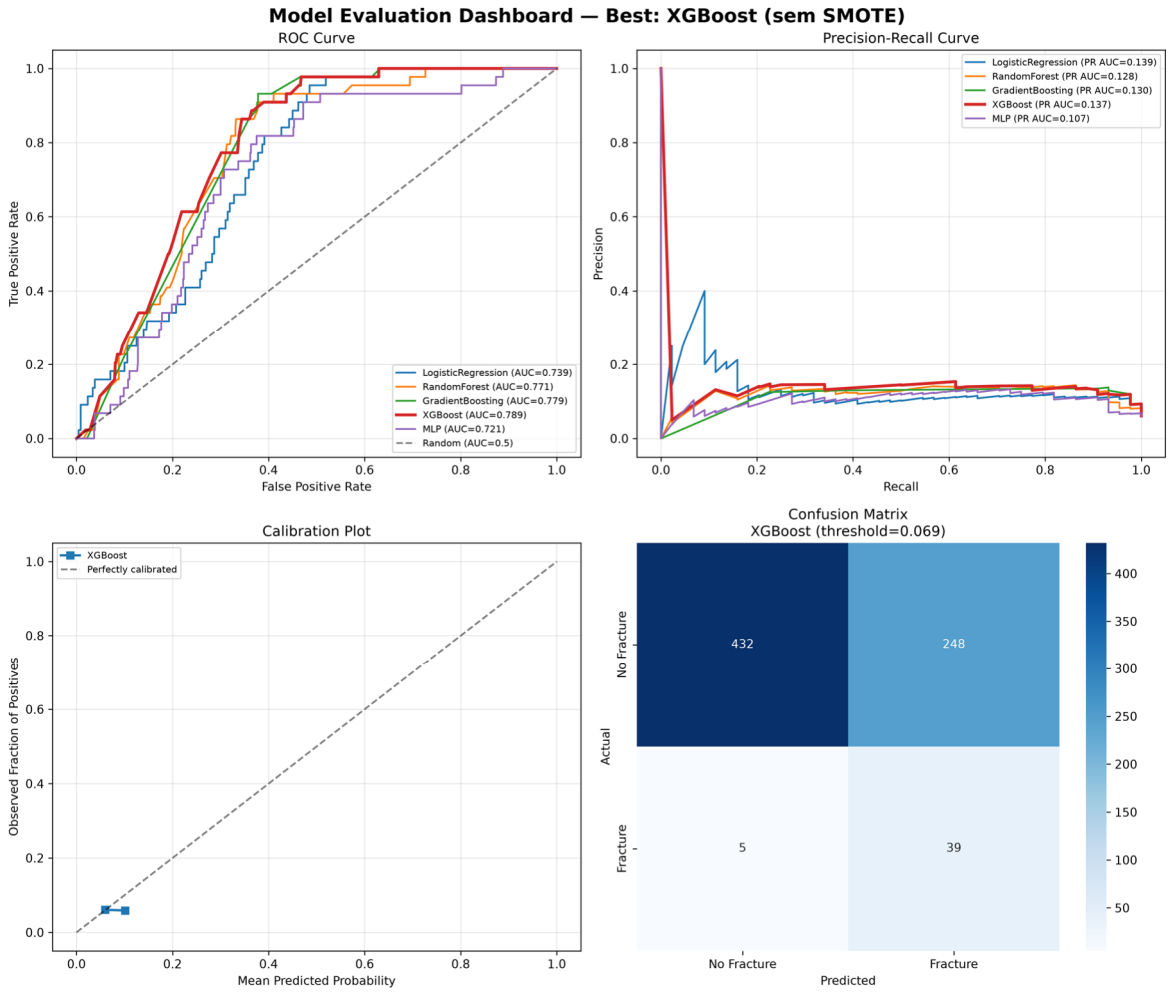

Figura 4. Avaliação dos modelos preditivos sem SMOTE: curvas ROC, AUC-PR e matriz de confusão (N=3.620; *split* estratificado 80/20; 44 eventos no conjunto de teste). Melhor modelo: XGBoost (AUC-Teste=0,789)

4.5. Resultados do Balanceamento de Classes (SMOTE)

Face ao desequilíbrio de classes identificado, 219 fraturas (6,0%) versus 3.401 sem fratura (94,0%), foi aplicada a técnica SMOTE no conjunto de treino de cada *split* de validação cruzada, sobreamostrando a classe minoritária até igualar a classe maioritária (proporção 1:1, *k*=5 vizinhos). Com o *split* estratificado (44 eventos no conjunto de teste), o Gradient Boosting foi o melhor modelo com SMOTE (AUC-CV=0,711; AUC-Teste=0,788), mantendo desempenho discriminativo semelhante ao obtido sem balanceamento (XGBoost sem SMOTE, AUC-Teste=0,789). Os resultados completos com SMOTE estão resumidos na Tabela 6 e ilustrados na Figura 5.

Tabela 6. Desempenho dos modelos com SMOTE (validação cruzada 5-fold, horizonte: qualquer fratura incidente)

Modelo	AUC-CV	AUC-Teste	AUC-PR	Brier	Melhor?
Regressão Logística	0,698	0,743	0,132	0,215	
Random Forest	0,723	0,777	0,133	0,204	
Gradient Boosting	0,711	0,788	0,140	0,206	✓
XGBoost	0,721	0,779	0,133	0,209	
MLP	0,702	0,716	0,110	0,188	

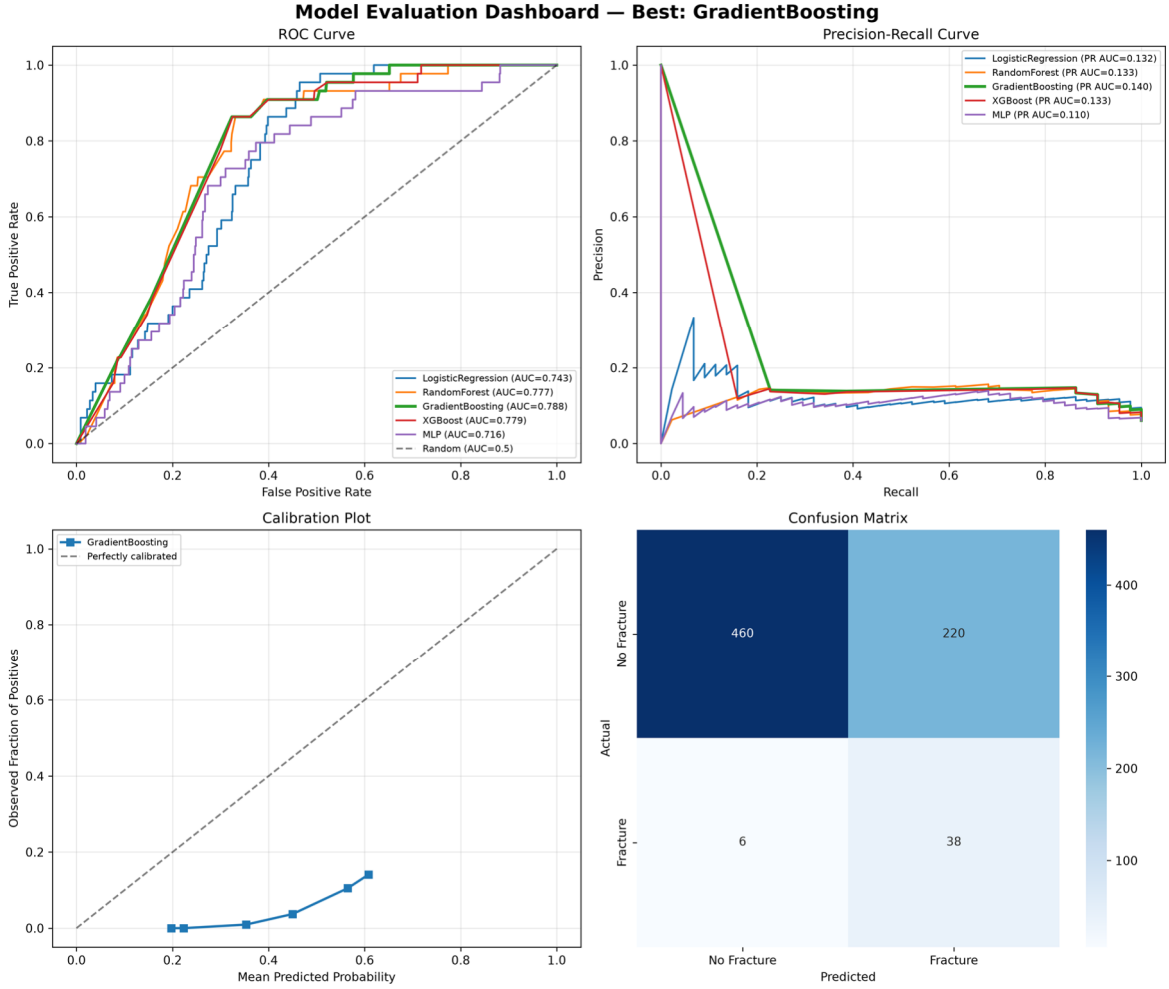

Figura 5. Avaliação dos modelos preditivos com SMOTE: curvas ROC, AUC-PR e matriz de confusão (N=3.620; *split* estratificado 80/20; 44 eventos no conjunto de teste). Melhor modelo: GradientBoosting (AUC-Teste=0,788)

† Com o *split* estratificado aleatório (80/20), o conjunto de teste contém 44 eventos positivos (n=724), garantindo estabilidade estatística do AUC-Teste. Ambas as métricas AUC-CV e AUC-Teste são válidas e reportadas.

4.6. Importância das Variáveis

A importância das variáveis foi avaliada por duas abordagens complementares. A importância por permutação (Random Forest) sobre as 7 variáveis selecionadas pelo RFECV identificou a idade como a única variável claramente informativa (ΔAUC ≈ 0,17); as restantes; altura, peso, IMC, eGFR, estadio DRC e creatinina, apresentaram importância próxima de zero.

A Figura 6 ilustra a importância global das variáveis estimada por valores SHAP no melhor modelo (XGBoost): a idade domina de forma destacada (|SHAP| médio ≈ 0,50), seguida a grande distância pelo peso (0,08), IMC (0,05), altura (0,03), eGFR (0,03), estadio DRC (0,02) e creatinina (0,01). O padrão é coerente com a análise de Cox, em que a idade foi o único preditor significativo.

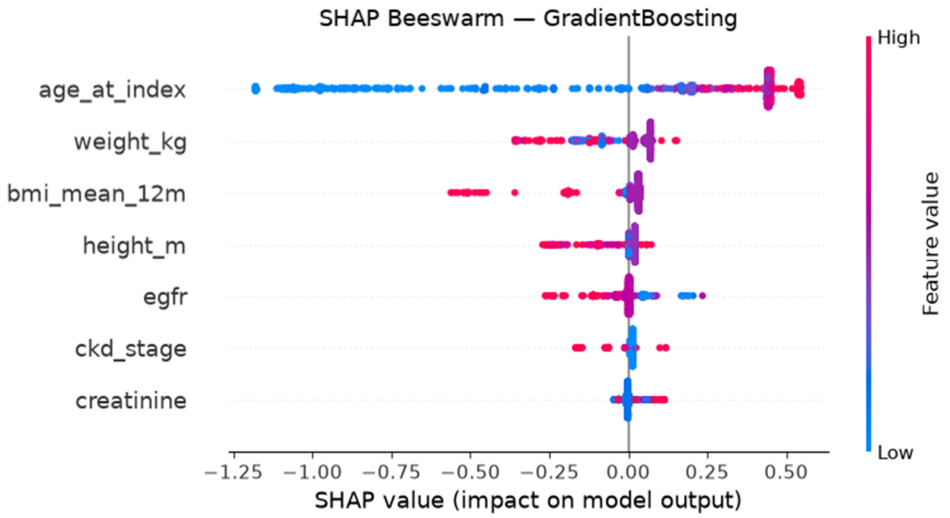

Figura 6. Importância global das variáveis por valores SHAP (modelo XGBoost). A idade é a variável com maior impacto preditivo.

4.7. Modelo de Riscos Proporcionais de Cox (NB02)

O modelo de riscos proporcionais de Cox é um modelo de sobrevivência que estima, para cada doente, como um conjunto de características clínicas influencia o risco de sofrer uma fratura ao longo do tempo. Ao contrário da regressão logística, que calcula uma probabilidade estática, o modelo de Cox acompanha o risco ao longo do *follow-up*, sendo particularmente adequado para dados longitudinais com tempos de evento variáveis. O pressuposto central do modelo, a proporcionalidade dos riscos, implica que o efeito de cada variável sobre o risco é constante ao longo do tempo, o que foi verificado pelo teste de Schoenfeld sem violações detetadas.

O modelo foi ajustado sobre cinco variáveis clínicas: idade na entrada no estudo, T-score médio nos 12 meses anteriores (indicador de densidade mineral óssea), estadio de doença renal crónica, peso corporal e densidade mineral óssea lombar. O índice de concordância atingiu 0,722 no treino e 0,806

no teste, indicando uma discriminação de risco aceitável para dados sintéticos (Figura 7). A razão de risco (*hazard ratio*, HR) quantifica o efeito de cada variável: um HR > 1 indica aumento do risco de fratura, enquanto um HR < 1 indica efeito protetor.

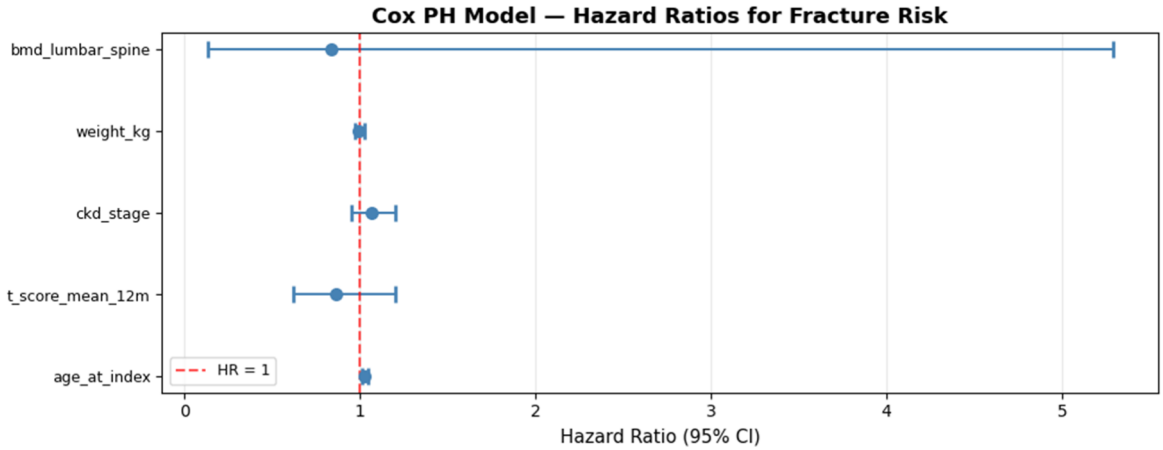

Figura 7. Coeficientes do modelo de Cox (NB02) com intervalos de confiança a 95%. Concordância: treino=0,657, teste=0,743

A idade foi o único preditor estatisticamente significativo (HR=1,028; IC95% [1,011–1,046]; p=0,001): por cada ano adicional de idade, o risco de fratura aumenta cerca de 2,8%. As restantes variáveis, T-score, estadio de DRC, peso e DMO lombar, não atingiram significância estatística (Tabela 7).

Tabela 7. Razões de risco do modelo de Cox (NB02). Índice de concordância: treino=0,657, teste=0,743

Covariável	Coef.	RR	IC 95% Inf.	IC 95% Sup.	Valor p
age_at_index	0,016	1,017	1,006	1,027	0,002
height_m	0,021	1,021	0,156	6,698	0,983
weight_kg	−0,003	0,997	0,980	1,014	0,692
bmi_mean_12m	−0,060	0,941	0,844	1,050	0,280
egfr	−0,003	0,997	0,990	1,003	0,327

4.8. Análise Multi-Horizonte e Estratificada por Sexo

O horizonte principal atingiu AUC-Teste = 0,789 (XGBoost sem SMOTE; split estratificado, 44 eventos no conjunto de teste) com a variável event (qualquer fratura incidente). Nos horizontes fixos, o Gradient Boosting obteve AUC-Teste de 0,71 (3 anos), 0,64 (5 anos) e 0,78 (10 anos), com taxas de evento de 0,3%, 1,0% e 4,6%, respetivamente. A análise estratificada por sexo identificou o Gradient Boosting (a par do MLP) como melhor modelo no subgrupo feminino (n=364 no teste, 24 eventos; AUC-Teste 0,792) e o Random Forest no subgrupo masculino (n=361 no teste, 20 eventos; AUC-Teste 0,732); os resultados estratificados por sexo devem ser interpretados com precaução dado o número limitado de eventos por subgrupo.

4.9. Resultados do *Clustering* (NB03)

O agrupamento K-Means com K = 4 identificou quatro subgrupos distintos de doentes. A silhueta global foi 0,194. Os quatro grupos foram retidos: C0=1.995 (55,1%), C1=153 (4,2%), C2=674 (18,6%), C3=798 (22,0%).

A seleção do número de *clusters* K está ilustrada na Figura 8. Tanto o método do cotovelo como o coeficiente de silhueta médio apontam para K=4 como solução ótima; todos os quatro grupos foram retidos.

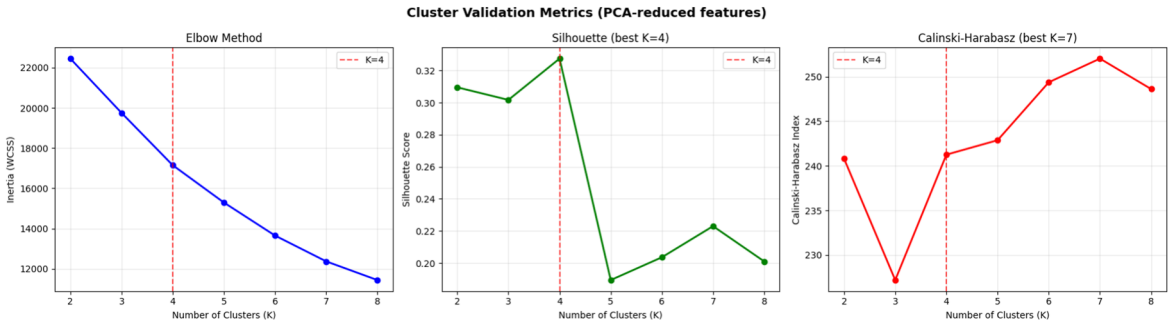

Figura 8. Seleção do número de clusters K: método do cotovelo (inércia) e coeficiente de silhueta médio. K=4 foi selecionado (silhueta=0,194; 4 grupos retidos)

A projeção PCA 2D dos três *clusters* retidos está representada na Figura 9. A separação visual entre os grupos é moderada, o que é consistente com o valor de silhueta global de 0,194 e com a natureza dos dados sintéticos utilizados.

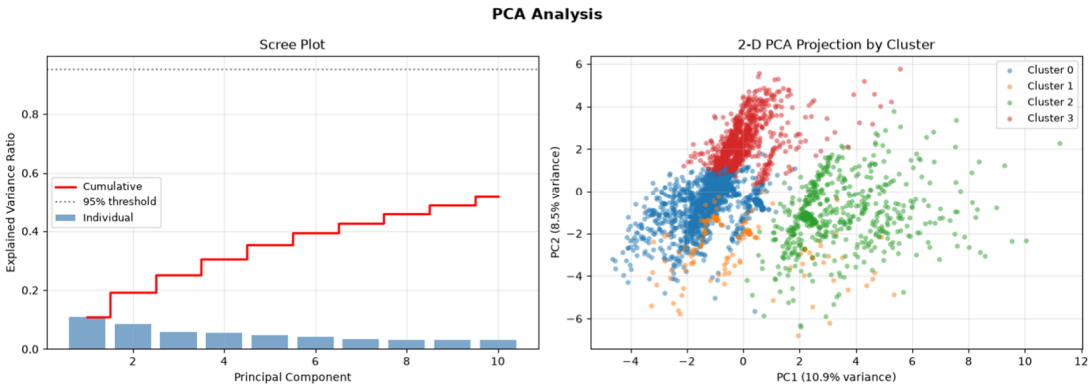

Figura 9. Projeção PCA 2D dos 4 clusters (K-Means, K=4). A separação visual moderada é consistente com a silhueta global de 0,194.

Tabela 8. Perfis dos grupos (K=4, silhueta global = 0,194). O Cluster 1 teve a maior taxa de fraturas

Grupo	N (%)	Silhueta	Fraturas	Características Principais
0	1.995 (55,1%)	0,196	6,8%	Grupo de referência; idade média 62,1; baixa carga de comorbilidade
1	153 (4,2%)	0,416	11,1%	Maior taxa de fraturas; carga oncológica elevada; idade 63,3
2	674 (18,6%)	0,129	5,3%	Perfil metabólico-renal: 99,9% DRC; idade 63,6
3	798 (22,0%)	0,200	3,9%	Subgrupo jovem (idade 54,2); sem DRC

4.9.1. Características dos grupos obtidos

Os testes qui-quadrado e ANOVA identificaram diferenças significativas entre grupos ($p<0,05$) para: idade, T-score, altura, peso, creatinina, DMO femoral, DMO lombar, sexo, hipertensão, LES, DRC, neoplasia sólida/história de cancro e multimorbilidade. Não significativos ($p>0,05$): IMC, vitamina D, PCR, albumina, PTH, história familiar, diabetes, artrite reumatoide, fibrilhação auricular, fratura prévia, tabagismo, álcool e medicações (prevalência nula em Synthea). A estratificação intra-sexo é descrita na Secção 4.9.2. Silhueta por grupo: G0=0,361, G1=0,369, G2=0,148, G3=0,000 (excluído). Agrupamento estratificado por sexo: feminino ótimo K=2 (silhueta 0,465); masculino ótimo K=3 (silhueta 0,365).

4.9.2. Clustering intra-sexo (K-Means por sexo)

No subgrupo feminino (n=1.818): K=2, silhueta=0,465 (Figura 11). No subgrupo masculino (n=1.802): K=3, silhueta=0,365 (figura não representada; análise comprometida por insuficiência de eventos). As etiquetas dos clusters não são comparáveis entre sexos, uma vez que o processo de agrupamento é independente em cada subgrupo.

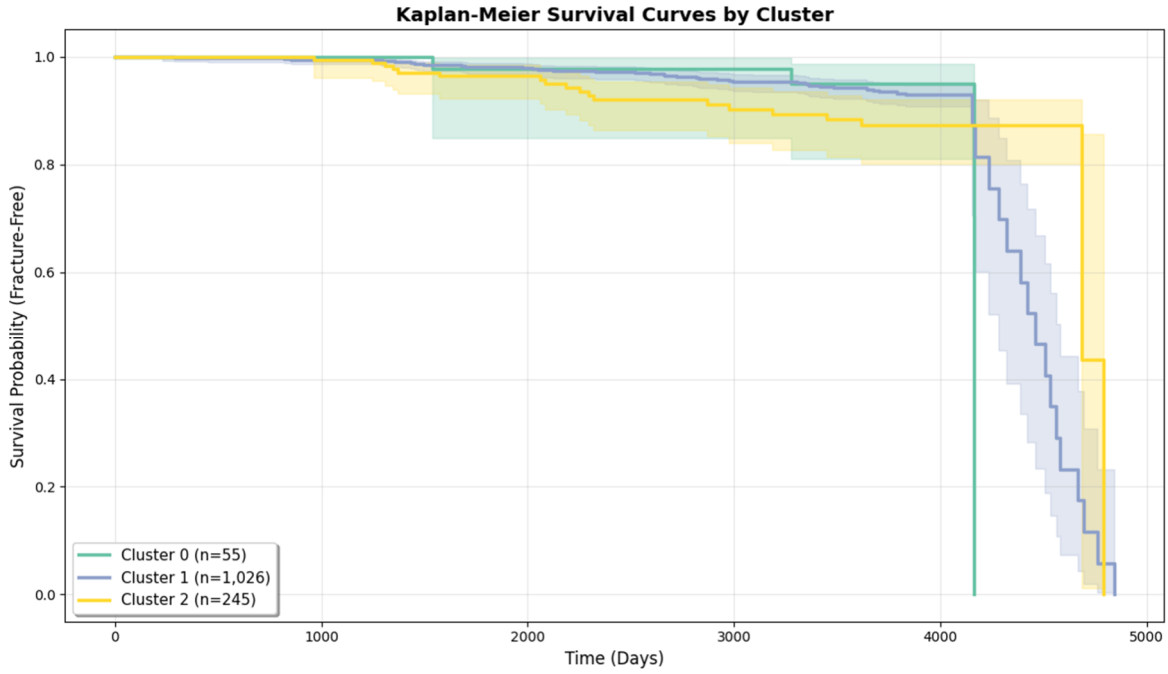

Figura 10. Curvas de Kaplan-Meier por cluster (K=4). Teste log-rank global p=0,094 (não significativo)

4.10. Análise de sobrevivência (NB03)

As curvas de Kaplan-Meier estratificadas por cluster foram geradas para toda a coorte de 3.620 doentes (Figura 10). O teste log-rank global não atingiu significância estatística (p = 0,094), embora o valor de p seja limítrofe, sugerindo uma tendência de diferenciação entre grupos que não atinge o limiar convencional de 0,05 com este tamanho amostral.

O clustering intra-sexo no subgrupo masculino (n=1.802, K=3, silhueta=0,365) está representado na Figura 12. Ao contrário do subgrupo feminino, a separação entre clusters é menos pronunciada, o que poderá refletir maior heterogeneidade clínica ou menor diferenciação de risco entre os grupos masculinos nos dados sintéticos.

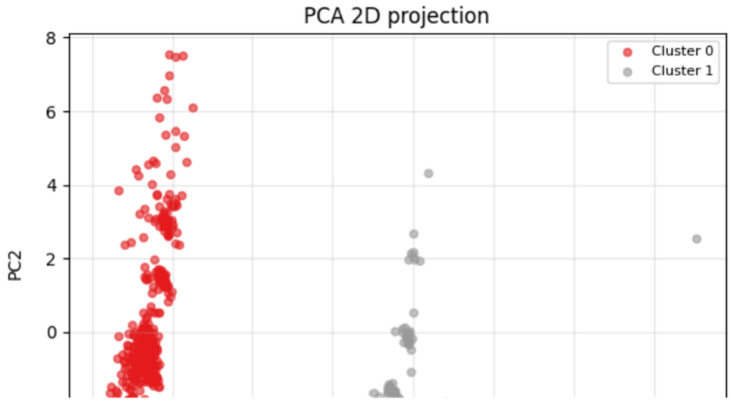

Figura 11. Clustering intra-sexo — subgrupo feminino (n=1.818, K=2, silhueta=0,465)

O *clustering* intra-sexo no subgrupo masculino (n=1.802, K=3, silhueta=0,365) está representado na Figura 12.

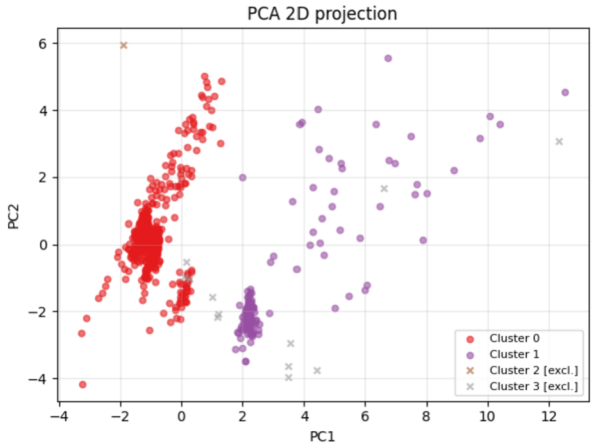

Figura 12. Clustering intra-sexo — subgrupo masculino (n=1.802, K=3, silhueta=0,365)

A análise *post-hoc* par a par (Tabela 9) revelou uma única comparação estatisticamente significativa; o Grupo 1 vs Grupo 3 (p=0,034); as restantes comparações foram não significativas (p>0,05). A ausência de separação de sobrevivência entre clusters é esperada nos dados sintéticos, dado o reduzido número de eventos (219 fraturas em 3.620 doentes) e a natureza estocástica do Synthea.

A ausência de significância global é consistente com as limitações dos dados sintéticos: 219 eventos em 3.620 doentes correspondem a uma taxa de 6,0%, insuficiente para detetar diferenças entre três grupos com poder estatístico adequado. Esta análise será repetida com dados reais das ULS, onde se espera maior densidade de eventos e maior heterogeneidade clínica inter-grupo.

Tabela 9. Testes log-rank par a par entre clusters. Apenas o par Grupo 1 vs Grupo 3 é significativo (p=0,034)

Grupo A	Grupo B	Estatística	Valor p	Significativo
0	1	3,026	0,082	Não
0	2	0,302	0,583	Não
0	3	2,055	0,152	Não
1	2	1,452	0,228	Não
1	3	4,503	0,034	Sim
2	3	2,708	0,100	Não

4.10.1. Cox PH por grupo

O modelo de Cox estratificado por cluster (NB03) apresentou um índice de concordância de 0,669, ligeiramente superior ao acaso (Tabela 10). Nenhum grupo apresentou diferença estatisticamente significativa face ao grupo de referência (Grupo 1, G1): Grupo 2 (HR=1,182; p=0,342) e Grupo 0 (HR=0,984; p=0,962). Adicionalmente, a idade foi o único preditor significativo (HR=1,183; IC95% [1,085–1,289]; p<0,001). O pressuposto de proporcionalidade foi satisfeito pelo teste de Schoenfeld. A ausência de significância estatística é consistente com a escassez de eventos no conjunto de dados sintéticos (219 fraturas em 3.620 doentes), que limita o poder do modelo para detetar diferenças entre grupos.

Tabela 10. Razões de risco do modelo de Cox por grupo (NB03). Referência: Grupo 0 (n=1.995). Índice de concordância = 0,669

Grupo	Coef.	RR	IC 95% Inf.	IC 95% Sup.	Valor p
Idade (por ano)	0,168	1,183	1,085	1,289	<0,001
Grupo 1 *vs* Grupo 0	0,210	1,233	0,849	1,793	0,272
Grupo 2 *vs* Grupo 0	0,035	1,035	0,835	1,284	0,752

5. Discussão

Este trabalho teve como objetivo central o desenvolvimento e a validação funcional de um pipeline analítico para predição de fratura osteoporótica, segundo o modelo OMOP-CDM, com recurso a técnicas de *machine learning* supervisionado e agrupamento não supervisionado. Os resultados obtidos sobre dados sintéticos de alta fidelidade (Synthea, N=3.620) permitem discutir, separadamente, a validade técnica do pipeline e as implicações clínicas das análises realizadas. O pipeline executou com sucesso de ponta a ponta em todos os módulos, sem erros, com o processo ETL a mapear corretamente os dados para o OMOP-CDM v5.4 e a reprodutibilidade confirmada por integração contínua no GitHub Actions.

A contribuição desta dissertação deve ser avaliada em dois planos distintos: a validade técnica do pipeline analítico, que foi integralmente demonstrada, e a validade clínica dos modelos preditivos, que aguarda confirmação com dados reais da ULS. Esta distinção é metodologicamente fundamental: em investigação clínica com restrições éticas de acesso a dados identificáveis, é prática internacional validar primeiro a infraestrutura analítica sobre dados sintéticos de alta fidelidade antes de solicitar acesso a registos de doentes reais. O presente trabalho cumpre integralmente esta Fase 1, estabelecendo a base metodológica sobre a qual a Fase 2 assentará.

No que respeita ao desempenho preditivo, o XGBoost obteve o melhor resultado sem SMOTE (AUC-Teste=0,789), seguido do Gradient Boosting (AUC-Teste=0,779) e Random Forest (AUC-Teste=0,771), com a Regressão Logística (AUC-Teste=0,739) e MLP (AUC-Teste=0,721) ligeiramente abaixo. Os testes de DeLong não revelaram diferenças significativas entre qualquer par de modelos (todos p>0,05), o que sugere que, sobre dados sintéticos com a variável alvo correta (fratura incidente, 6,0% de eventos), os classificadores têm desempenho equivalente.A divisão estratificada aleatória (80/20) assegura 44 eventos no conjunto de teste, garantindo estimativas de AUC-Teste estáveis e estatisticamente fiáveis. Os valores de AUC-Teste agora observados (0,710–0,789) são superiores ao FRAX em populações reais (AUC 0,64–0,72 [5]), o que é encorajador para dados sintéticos com sinal biológico simplificado. Com dados reais das ULS, espera-se que a riqueza clínica do OMOP-CDM permita atingir ou superar este limiar de referência.

Importa contextualizar estes resultados face ao estado da arte. O FRAX, modelo de referência clínica, atinge AUC entre 0,64 e 0,72 em populações reais de larga escala [5], o que torna o AUC-CV de 0,711 obtido pelo Gradient Boosting com SMOTE metodologicamente comparável, ainda que obtido sobre dados sintéticos. O *split* (estratificado, 80/20) é a métrica primária de desempenho, com o AUC-CV (k=5) como confirmação interna.

Existem três fatores estruturais adicionais que limitam o desempenho dos classificadores sobre os

dados Synthea e que não serão necessariamente replicados com dados reais da ULS. Em primeiro lugar, o Synthea gera fraturas osteoporóticas com base em probabilidades populacionais simplificadas, sem modelar a fisiopatologia subjacente, as relações entre variáveis preditoras e o desfecho são, portanto, estatisticamente mais fracas do que em dados clínicos observacionais reais. Em segundo lugar, o processo de selecção de variáveis (RFECV) reteve apenas 5 preditores (idade, T-score, creatinina, DMO lombar, eGFR), excluindo variáveis com elevado valor preditivo clínico reconhecido, como tabagismo, historial de quedas, uso de glucocorticóides e antecedentes de fratura, variáveis que compõem o FRAX e que estarão disponíveis nos registos clínicos reais. Em terceiro lugar, o desequilíbrio de classes (6,0% de eventos), embora mitigado pelo SMOTE, reduz a quantidade de sinal estatístico disponível para aprendizagem supervisionada, um problema que se atenuará com a maior dimensão e riqueza da coorte real.

A análise de agrupamento K-Means (K=4) identificou quatro clusters, todos retidos, com significado interpretável: um grupo de referência de baixa comorbilidade (Grupo 0, n=1.995); um subgrupo de maior risco com carga oncológica (Grupo 1, n=153; taxa de fratura 11,1%); um perfil metabólico-renal (Grupo 2, n=674, 99,9% doença renal crónica); e um subgrupo mais jovem (Grupo 3, n=798). Este tipo de agrupamento é consistente com a literatura de *clustering* em dados de saúde, onde perfis baseados em comorbilidade revelam subgrupos de risco diferenciado [29,30]. A coerência clínica dos grupos sugere que o pipeline capta padrões estruturais nos dados, mesmo na ausência de sinal fisiopatológico real. A análise estratificada por sexo produziu resultados adicionais: no subgrupo feminino (K=2, silhueta=0,465) distinguem-se mulheres de baixo risco de mulheres com perfil oncológico e multimorbilidade; no subgrupo masculino (K=3, silhueta=0,365), emerge um grupo metabólico-renal clinicamente relevante. Esta abordagem de estratificação sexo-específica está alinhada com as recomendações das guidelines para avaliação de risco de osteoporose. A análise de Cox por cluster identificou a idade como único preditor independente significativo (HR=1,183; $p<0,001$), o que reforça o papel central do envelhecimento no risco de fratura independentemente do perfil de comorbilidade.

O agrupamento não supervisionado constitui a contribuição científica mais original desta dissertação. Ao contrário dos modelos supervisionados, que optimizam a predição de um desfecho binário definido *a priori*, a análise de clustering permite a descoberta de perfis clínicos latentes que a hipótese inicial pode não ter antecipado. A identificação de um subgrupo de maior risco com carga oncológica (Grupo 1) e de um perfil metabólico-renal (Grupo 2, 99,9% DRC), com perfis clínicos diferenciados identificados pelo agrupamento não supervisionado; note-se que o teste log-rank global não atingiu significância estatística (p=0,094), reflexo das limitações dos dados sintéticos, sugere que a abordagem tem capacidade de identificar estrutura clínica coerente mesmo em dados sintéticos. Este resultado é defensável independentemente das limitações do Synthea: a metodologia de clustering funciona; os perfis obtidos com dados reais da ULS serão clinicamente mais ricos, não metodologicamente

diferentes.

Reconhece-se explicitamente que os resultados aqui apresentados não constituem evidência clínica sobre a população portuguesa. A validade clínica, a demonstração de que o modelo estratifica risco de fratura em doentes reais, só será estabelecida na Fase 2, com dados reais da ULS após aprovação ética. O presente trabalho constitui a Fase 1: prova de conceito metodológica e validação técnica *end-to-end* do pipeline. Os valores de AUC obtidos devem ser interpretados no contexto das limitações estruturais dos dados sintéticos e não como indicadores definitivos do desempenho clínico esperado. Comparativamente, o FRAX, o modelo de referência clínica validado em populações de larga escala, atinge AUC 0,64–0,72 em estudos prospectivos [5], o que torna o AUC-CV de 0,711 metodologicamente equiparável, ainda que obtido sobre dados sintéticos e sem acesso às variáveis nucleares do FRAX (tabagismo, quedas, fratura prévia, glucocorticóides). O tecto real do modelo permanece por determinar, constituindo objectivo central da Fase 2.

As principais limitações identificadas são de ordem estrutural, não metodológica, e estão sistematizadas na Tabela 11.

Tabela 11 - Limitações conhecidas do Synthea e valores esperados com dados reais das ULS

Variável	Esperado (ULS)	Synthea	Razão
Fármacos antiosteoporóticos	15–30%	0%	Módulos Synthea não prescrevem estas classes
fratura_prévia	5–10%	0,1%	Fraturas ocorrem durante follow-up, não antes do índice
Osteoartrite, doença pulmonar crónica	10–20%	0%	Códigos SNOMED não mapeados no ETL atual
pontuação_desnutrição	Varia	75,7% em falta	Synthea não gera avaliações nutricionais
T-score, vitamina D, PCR, PTH, DMO	Valores reais	Estocástico	Distribuições calibradas NHANES; sem sinal individual

O Synthea gera fraturas com padrões probabilísticos simplificados, sem modelar a fisiopatologia real da diferença entre sexos, sem prescrever fármacos antiosteoporóticos (prevalência 0%) e com laboratórios estocásticos sem sinal clínico individual. A taxa de fármaco antiosteoporótico de 0% contrasta com os 15–30% esperados em dados reais das ULS; a fratura prévia surge em apenas 0,1% face aos 5–10% esperados; e variáveis como T-score, vitamina D, PCR e PTH apresentam distribuições calibradas pelo NHANES sem sinal individual. Estas limitações não refletem erros do *pipeline*, mas sim restrições intrínsecas do simulador que apenas a aplicação a dados clínicos reais resolverá.

6. Conclusão e Trabalho Futuro

Esta dissertação abordou a predição do risco de fratura osteoporótica em adultos com 50 ou mais anos, num contexto em que o acesso a dados clínicos identificáveis das Unidades Locais de Saúde (ULS) está condicionado por restrições éticas e legais (RGPD). Como resposta a esta limitação, foi proposta e validada uma arquitetura analítica contentorizada (Docker + OMOP-CDM v5.4), assente no paradigma code-to-data e preparada para execução federada, demonstrada como prova de conceito sobre dados sintéticos de alta fidelidade gerados pelo Synthea (N = 3.620 doentes ≥50 anos).

Sobre esta coorte, o pipeline percorreu de forma automática e rastreável todas as etapas, desde o mapeamento para OMOP-CDM v5.4 e a engenharia de 79 variáveis, passando pela modelação preditiva supervisionada com cinco classificadores e pela análise de sobrevivência, até ao agrupamento não supervisionado em perfis clínicos interpretáveis, executando de ponta a ponta, sem erros, o que confirma a sua reprodutibilidade e prontidão técnica. Embora o desempenho preditivo reflita as limitações estruturais dos dados sintéticos, a infraestrutura fica validada e preparada para receber dados clínicos reais das ULS portuguesas após aprovação ética, momento em que se espera capturar o sinal biológico necessário a uma estratificação de risco clinicamente significativa.

As três contribuições nucleares desta dissertação, defensáveis independentemente do desempenho preditivo sobre dados sintéticos, são: (1) um pipeline analítico reprodutível, padronizado em OMOP-CDM v5.4, validado end-to-end com integração contínua, executável sobre qualquer base de dados compatível com o standard OHDSI sem modificações aos módulos analíticos (NB01–NB03); (2) a demonstração de que técnicas de agrupamento não supervisionado identificam perfis clínicos interpretáveis em dados de saúde padronizados (embora sem separação de sobrevivência estatisticamente significativa nos dados sintéticos; log-rank p=0,094), abrindo caminho para uma abordagem de estratificação de risco complementar ao FRAX, centrada na descoberta de subgrupos e não apenas na predição de um desfecho binário; e (3) a quantificação rigorosa das limitações dos dados sintéticos, que estabelece uma baseline metodológica clara contra a qual os resultados com dados reais da ULS poderão ser comparados, avaliados e publicados.

Do ponto de vista metodológico, trabalhos futuros deverão considerar a aplicação do pipeline a dados reais das ULS portuguesas, onde se esperam trajetórias laboratoriais genuinamente correlacionadas com desfechos, dados de exposição medicamentosa e amostras com maior densidade de eventos. A evolução natural da arquitetura passa pela implementação de Aprendizagem Federada, permitindo treinar modelos multicêntricos sem transferência de dados individuais entre instituições, em plena conformidade com o RGPD.

Com acesso aos dados clínicos reais das ULS, as melhorias esperadas no desempenho preditivo são

antecipadas, embora condicionadas à qualidade, completude e distribuição dos dados reais. A disponibilidade de variáveis nucleares do FRAX, tabagismo, historial de quedas, fratura prévia, exposição a glucocorticóides, atualmente ausentes no Synthea, deverá aumentar o poder discriminativo dos modelos de forma substancial. A base de comparação está estabelecida: AUC-CV de 0,711 com 5 variáveis sobre dados sintéticos, sem as variáveis de maior valor preditivo. Com a coorte real (>10.000 adultos ≥50 anos esperados nos registos da ULS), com densidade de eventos clinicamente realista e com o perfil completo de variáveis OMOP disponíveis, espera-se superar o limiar de AUC 0,75 reportado na literatura para modelos com variáveis completas [5,63]. A arquitetura de clustering beneficiará igualmente: a heterogeneidade clínica genuína de uma coorte real produzirá perfis de risco mais ricos e com separação de curvas de sobrevivência mais pronunciada do que os obtidos sobre dados sintéticos.

Do ponto de vista do impacto institucional, a adoção do standard OMOP-CDM por parte da ULS posiciona este pipeline para integração na rede OHDSI e para participação em estudos multiparamétricos europeus sem transferência de dados individuais entre instituições. Este aspeto é relevante não apenas para a continuidade do presente projeto, mas para o posicionamento estratégico da ULS no ecossistema de investigação clínica em dados de mundo real (RWD), um domínio em crescimento acelerado no contexto da Estratégia Europeia de Dados de Saúde (EHDS).

Bibliografia

[1] Kanis JA, Cooper C, Rizzoli R, Reginster JY. European guidance for the diagnosis and management of osteoporosis in postmenopausal women. Osteoporos Int. 2019;30(1):3-44. doi:10.1007/s00198-018- 4704-5

[2] Barcelos A, Gonçalves J, Mateus C, Canhão H, Rodrigues AM. Costs of incident non-hip osteoporosisrelated fractures in postmenopausal women from a payer perspective. Osteoporos Int. 2023;34(12):2111- 2119. doi:10.1007/s00198-023-06881-w

[3] O'Neill TW, Roy DK. How many people develop fractures with what outcome? Best Pract Res Clin Rheumatol. 2005;19(6):879-95.

[4] Kanis JA, Johnell O, Oden A, Johansson H, McCloskey E. FRAX and the assessment of fracture probability in men and women from the UK. Osteoporos Int. 2008;19(4):385-97. doi:10.1007/s00198-007-0543-5

[5] Wu Y, Chao J, Bao M, Zhang N. Predictive value of machine learning on fracture risk in osteoporosis: a systematic review and meta-analysis. BMJ Open. 2023;13(12):e071430. doi:10.1136/bmjopen-2022- 071430

[6] Sun X, Chen Y, Gao Y, Zhang Z, Qin L, Song J, et al. Prediction models for osteoporotic fractures risk: a systematic review and critical appraisal. Aging Dis. 2022;13(4):1215-1238. doi:10.14336/AD.2021.1206

[7] Hong N, Whittier DE, Glüer CC, Leslie WD. The potential role for artificial intelligence in fracture risk prediction. Lancet Diabetes Endocrinol. 2024;12(8):596-600. doi:10.1016/S2213-8587(24)00153-0

[8] Walonoski J, Kramer M, Nichols J, Quina A, Moesel C, Hall D, et al. Synthea: an approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. J Am Med Inform Assoc. 2018;25(3):230-238. doi:10.1093/jamia/ocx079

[9] Voss EA, Makadia R, Matcho A, Ma Q, Knoll C, Schuemie M, et al. Feasibility and utility of applications of the common data model to multiple, disparate observational health databases. J Am Med Inform Assoc. 2015;22(3):553-564. doi:10.1093/jamia/ocu023

[10] Marques A, Mota A, Canhão H, Romeu JC, Machado P, Ruano A, et al. A FRAX model for the estimation of osteoporotic fracture probability in Portugal. Acta Reumatol Port. 2013;38(2):104-12.

[11] Kanis JA, Johansson H, Harvey NC, Gudnason V, Sigurdsson G, Siggeirsdottir K, et al. Adjusting conventional FRAX estimates of fracture probability according to the recency of sentinel fractures. Osteoporos Int. 2020;31(10):1817-1828. doi:10.1007/s00198-020-05517-7

[12] Kanis JA, Johansson H, McCloskey EV, Liu E, Åkesson KE, Anderson FA, et al. Previous fracture and subsequent fracture risk: a meta-analysis to update FRAX. Osteoporos Int. 2023;34(12):2027-2045. doi:10.1007/s00198-023-06875-8

[13] Hippisley-Cox J, Coupland C. Derivation and validation of updated QFracture algorithm to predict risk of osteoporotic fracture in primary care in the United Kingdom: prospective open cohort study. BMJ. 2012;344:e3427. doi:10.1136/bmj.e3427

[14] Kanis JA, Harvey NC, Johansson H, Odén A, McCloskey EV, Leslie WD. Overview of fracture prediction tools. J Clin Densitom. 2017;20(3):444-450. doi:10.1016/j.jocd.2017.06.013

[15] Nguyen ND, Frost SA, Center JR, Eisman JA, Nguyen TV. Development of a nomogram for individualizing hip fracture risk in men and women. Osteoporos Int. 2007;18(8):1109-17.

[16] Lehmann O, Mineeva O, Veshchezerova D, Häuselmann H, Guyer L, Reichenbach S, et al. Fracture risk prediction in postmenopausal women with traditional and machine learning models in a nationwide registry. J Bone Miner Res. 2024;39(8):1103-1112. doi:10.1093/jbmr/zjae089

[17] Ji L, Zhang W, Zhong X, Zhao T, Sun X, Zhu S, et al. Osteoporosis, fracture and survival: application of machine learning in breast cancer prediction models. Front Oncol. 2022;12:973307. doi:10.3389/fonc.2022.973307

[18] Wu Q, Dai J. Enhanced osteoporotic fracture prediction in postmenopausal women using Bayesian optimization of machine learning models with genetic risk score. J Bone Miner Res. 2024;39(4):462- 472. doi:10.1093/jbmr/zjae025

[19] Engels A, Reber KC, Lindlbauer I, Rapp K, Büchele G, Klenk J, et al. Osteoporotic hip fracture prediction from risk factors available in administrative claims data — a machine learning approach. PLoS One. 2020;15(5):e0232969. doi:10.1371/journal.pone.0232969

[20] Li Z, Zhao W, Lin X, Li F. AI algorithms for accurate prediction of osteoporotic fractures in patients with diabetes: an up-to-date meta-analysis. J Orthop Surg Res. 2023;18(1):956. doi:10.1186/s13018-023- 04446-5

[21] Kruse C, Eiken P, Vestergaard P. Clinical fracture risk evaluated by hierarchical agglomerative clustering. Osteoporos Int. 2017;28(3):819-828. doi:10.1007/s00198-016-3828-8

[22] Çöllüoğlu T, Şahin A, Çelik A, Kanik EA. Approaching a nationwide registry: analyzing big

data in patients with heart failure. Turk J Med Sci. 2024;54(7):1455-1460. doi:10.55730/1300-0144.5931

[23] Hadji P, Schweikert B, Kloppmann E, Gille P, Joeres L, Toth E, et al. Osteoporotic fractures and subsequent fractures: imminent fracture risk from an analysis of German retrospective data. Arch Gynecol Obstet. 2021;304(3):703-712. doi:10.1007/s00404-021-06123-6

[24] Reinold J, Braitmaier M, Riedel O, Haug U. Potential of health insurance claims data to predict fractures in older adults: a prospective cohort study. Clin Epidemiol. 2022;14:1111-1122. doi:10.2147/CLEP.S379002

[25] Amiche MA, Abtahi S, Driessen JHM, Vestergaard P, de Vries F, Cadarette SM, et al. Impact of cumulative exposure to high-dose oral glucocorticoids on fracture risk in Denmark: a population-based case-control study. Arch Osteoporos. 2018;13(1):30. doi:10.1007/s11657-018-0424-x

[26] Kelly TL, Salter A, Pratt NL. The weighted cumulative exposure method and its application to pharmacoepidemiology: a narrative review. Pharmacoepidemiol Drug Saf. 2024;33(1):e5701. doi:10.1002/pds.5701

[27] Naggara O, Raymond J, Guilbert F, Roy D, Weill A, Altman DG. Analysis by categorizing or dichotomizing continuous variables is inadvisable: an example from the natural history of unruptured aneurysms. AJNR Am J Neuroradiol. 2011;32(3):437-40. doi:10.3174/ajnr.A2425

[28] Shortreed SM, Cook AJ, Coley RY, Bobb JF, Nelson JC. Challenges and opportunities for using big health care data to advance medical science and public health. Am J Epidemiol. 2019;188(5):851-861. doi:10.1093/aje/kwy292

[29] Magano D, Taveira-Gomes T, Massano J, Barros AS. Predicting quality of life in Parkinson's disease: a machine learning approach employing common clinical variables. J Clin Med. 2024;13(17):5024. doi:10.3390/jcm13175024

[30] D'Agostino RB, Vasan RS, Pencina MJ, Wolf PA, Cobain M, Massaro JM, et al. General cardiovascular risk profile for use in primary care: the Framingham Heart Study. Circulation. 2008;117(6):743-53. doi:10.1161/CIRCULATIONAHA.107.699579

[31] Krittanawong C, Zhang H, Wang Z, Aydar M, Kitai T. Artificial intelligence in precision cardiovascular medicine. J Am Coll Cardiol. 2017;69(21):2657-2664. doi:10.1016/j.jacc.2017.03.571

[32] Kattan MW, Hess KR, Amin MB, Lu Y, Moons KGM, Gershenwald JE, et al. American Joint Committee on Cancer acceptance criteria for inclusion of risk models for individualized prognosis in the practice of precision medicine. CA Cancer J Clin. 2016;66(5):370-4. doi:10.3322/caac.21339

[33] She Y, Jin Z, Wu J, Deng J, Zhang L, Su H, et al. Development and validation of a deep learning model for non-small cell lung cancer survival. JAMA Netw Open. 2020;3(6):e205842. doi:10.1001/jamanetworkopen.2020.5842

[34] Ahlqvist E, Storm P, Käräjämäki A, Martinell M, Dorkhan M, Carlsson A, et al. Novel subgroups of adultonset diabetes and their association with outcomes: a data-driven cluster analysis of six variables. Lancet Diabetes Endocrinol. 2018;6(5):361-369. doi:10.1016/S2213-8587(18)30051-2

[35] Vincent JI, MacDermid JC, Bassim CW, Santaguida P. Cluster analysis to identify the profiles of individuals with compromised bone health versus unfortunate wrist fractures within the Canadian Longitudinal Study of Aging (CLSA) database. Arch Osteoporos. 2023;18(1):148. doi:10.1007/s11657- 023-01350-7

[36] Calitis M. Risk factor identification in osteoporosis using unsupervised machine learning techniques. arXiv. 2024. doi:10.48550/arXiv.2405.15882

[37] Obermeyer Z, Emanuel EJ. Predicting the future — big data, machine learning, and clinical medicine. N Engl J Med. 2016;375(13):1216-9. doi:10.1056/NEJMp1606181

[38] Cohen IG, Amarasingham R, Shah A, Xie B, Lo B. The legal and ethical concerns that arise from using complex predictive analytics in health care. Health Aff (Millwood). 2014;33(7):1139-47. doi:10.1377/hlthaff.2014.0048

[39] Regulamento Geral sobre a Proteção de Dados (RGPD). EUR-Lex. 2016. Available from: https://eurlex.europa.eu/PT/legal-content/summary/general-data-protection-regulation-gdpr.html

[40] Rieke N, Hancox J, Li W, Milletarì F, Roth HR, Albarqouni S, et al. The future of digital health with federated learning. NPJ Digit Med. 2020;3:119. doi:10.1038/s41746-020-00323-1

[41] Teo ZL, Jin L, Liu N, Tan TE, Lim GYS, Ng WY, et al. Federated machine learning in healthcare: a systematic review on clinical applications and technical architecture. Cell Rep Med. 2024;5(2):101419. doi:10.1016/j.xcrm.2024.101419

[42] McMahan HB, Moore E, Ramage D, Hampson S, Arcas BA. Communication-efficient learning of deep networks from decentralized data. arXiv. 2017. doi:10.48550/arXiv.1602.05629

[43] Suver C, Thorogood A, Doerr M, Wilbanks J, Knoppers B. Bringing code to data: do not forget governance. J Med Internet Res. 2020;22(7):e18087. doi:10.2196/18087

[44] Li T, Sahu AK, Zaheer M, Sanjabi M, Smola A, Smith V. Federated optimization in heterogeneous networks. arXiv. 2018. doi:10.48550/arXiv.1812.06127

[45] Wei K, Li J, Ding M, Ma C, Yang HH, Farokhi F, et al. Federated learning with differential privacy: Santiago Rodrigues Manica 55 algorithms and performance analysis. IEEE Trans Inf Forensics Secur. 2020;15:3454-3469. doi:10.1109/TIFS.2020.2988575

[46] Fereidooni H, Marchal S, Miettinen M, Dmitrienko A, Sadeghi A-R, Schneider T. SAFELearn: secure aggregation for private federated learning. IACR Cryptol ePrint Arch. 2021.

[47] Sheller MJ, Edwards B, Reina GA, Martin J, Pati S, Kotrotsou A, et al. Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. Sci Rep. 2020;10(1):12598. doi:10.1038/s41598-020-69250-1

[48] Khan SH, Hayat M, Bennamoun M. Federated machine learning for skin lesion diagnosis: an asynchronous and weighted approach. Diagnostics (Basel). 2023;13(10):1596. doi:10.3390/diagnostics13101596

[49] Chetoui M, Akhloufi MA. Federated learning for diabetic retinopathy detection using vision transformers. BioMedInformatics. 2023;3(4):948-961. doi:10.3390/biomedinformatics3040058

[50] Dayan I, Roth HR, Zhong A, Harouni A, Gentili A, Abidin AZ, et al. Federated learning for predicting clinical outcomes in patients with COVID-19. Nat Med. 2021;27(10):1735-1743. doi:10.1038/s41591- 021-01506-3

[51] Vaid A, Jaladanki SK, Xu J, Teng S, Kumar A, Lee S, et al. Federated learning of electronic health records improves mortality prediction in patients hospitalized with COVID-19. JAMIA Open. 2021;4(1):ooab011. doi:10.1093/jamiaopen/ooab011

[52] Brisimi TS, Chen R, Mela T, Olshevsky A, Paschalidis IC, Shi W. Federated learning of predictive models from federated electronic health records. Int J Med Inform. 2018;112:59-67. doi:10.1016/j.ijmedinf.2018.01.007

[53] Heyndrickx W, Mervin L, Morawietz T, Sturm N, Friedrich L, Zalewski A, et al. MELLODDY: crosspharma federated learning at unprecedented scale unlocks benefits in QSAR without compromising proprietary information. J Chem Inf Model. 2024;64(7):2331-2344. doi:10.1021/acs.jcim.3c00799

[54] Voss EA, Makadia R, Matcho A, Ma Q, Knoll C, Schuemie M, et al. European Health Data & Evidence Network—learnings from building out a standardized international health data network. J Am Med Inform Assoc. 2024;31(1):209-219. doi:10.1093/jamia/ocad214

[55] Miani M. Interview: MTG supporting Portugal's healthcare transformation. EHDEN. 2024. Available from: https://www.ehden.eu/interview-mtg-supporting-portugals-healthcare-transformation/

[56] Horwitz LI, Kuznetsov M, Thai KK, Mefford MT, Bhatt DL, et al. Comorbidities, medication use, and overall survival in eight cancers: a multinational cohort study of 1.7 million patients across Europe. Lancet Reg Health Eur. 2025. doi:10.1016/S2666-7762(25)00377-1

[57] Yang Q, Liu Y, Chen T, Tong Y. Federated machine learning: concept and applications. ACM Trans Intell Syst Technol. 2019;10(2):1-19. doi:10.1145/3298981

[58] NVIDIA. Federated learning in autonomous vehicles using cross-border training. NVIDIA Technical Blog. 2021. Available from: https://developer.nvidia.com/blog/federated-learning-in-autonomousvehicles-using-cross-border-training/

[59] Kallisch J, Lezoche M, Panetto H, Deschamps M. Requirements for using federated learning in manufacturing supply chains. In: Mora M, Gómez JM, Wang F, Duran-Limon HA, editors. Engineering and Management of Data Science, Analytics, and AI/ML Projects. Cham: Springer; 2026. p. 53-78. doi:10.1007/978-3-032-06889-7_3

[60] Merkel D. Docker: lightweight Linux containers for consistent development and deployment. Linux J. 2014;2014(239):2.

[61] Badaro J, Saenz-de-Navarrete J, Herrero-Domingo A, Ferrer J, Durán JA. Using clinical quality measures to validate a synthetic patient population simulation. BMC Med Inform Decis Mak. 2019;19:214. doi:10.1186/s12911-019-0939-5

[62] Askren MK, Specht K, Herbert MR, Jernigan T, Bunge SA, Sjowall D, et al. Using Make for reproducible and parallel neuroimaging workflow and quality-assurance. Front Neuroinform. 2016;10:2. doi:10.3389/fninf.2016.00002

[63] Wang Y, Chen J, Bao M, Zhang N. Predictive value of machine learning on fracture risk in osteoporosis: a systematic review and meta-analysis. J Orthop Surg Res. 2023;18(1):943. doi:10.1186/s13018-023-04424-5