

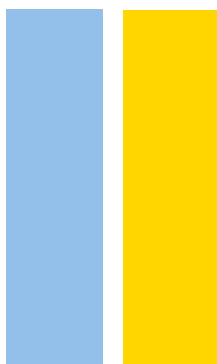
MESTRADO
INFORMÁTICA MÉDICA

Impacto dos Modelos Preditivos nas Infeções do Local Cirúrgico em Cirurgias de Coluna Lombar: Uma Revisão Sistemática

Rafaela Ventura

M

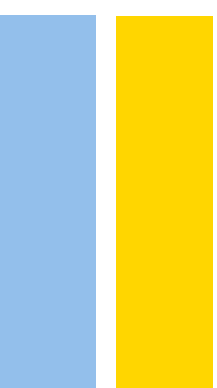
2025



SEDE ADMINISTRATIVA

FACULDADE DE **MEDICINA**

FACULDADE DE **CIÊNCIAS**



Agradecimentos

A conclusão desta dissertação não teria sido possível sem o apoio e contribuição de várias pessoas, às quais expresso o meu mais sincero reconhecimento.

À Professora Inês Dutra, manifesto a minha gratidão pela orientação científica atenta e pelo acompanhamento sempre disponível. A sua motivação constante foi essencial para que eu conseguisse superar as dificuldades encontradas ao longo deste percurso e concretizar esta tese.

Pelo companheirismo, motivação constantes e pela valiosa colaboração, agradeço à Daniela. A tua dedicação e apoio foram essenciais.

Aos meus pais e irmão, agradeço pelo apoio incondicional e pela insistência na conclusão deste capítulo da minha vida académica. O vosso apoio e confiança foram pilares fundamentais nesta jornada.

Por fim ao André, por me ter acompanhado e motivado ao longo de todo este processo, mesmo nas fases mais críticas e difíceis. A tua paciência, compreensão e apoio incansável foram determinantes para que conseguisse chegar até aqui.

A todos, o meu profundo obrigado.

Abstract

Context: Surgical site infections after lumbar spine surgery represent a significant complication with substantial clinical and economic impact. Predictive models have emerged as potential tools for risk stratification, but their methodological quality and clinical applicability remain unclear.

Methods: A systematic review was conducted following PRISMA guidelines with a protocol registered in PROSPERO (CRD42025613496). We searched PubMed, SCOPUS, EBSCO, ScienceDirect, and Cochrane Library for studies that developed or validated predictive models for SSI in lumbar spine surgery. Methodological quality was assessed using the PROBAST+AI tool for prediction model risk of bias assessment. Data extraction included study characteristics, model performance metrics, and predictor variables.

Results: Twelve studies were included, comprising 8 logistic regression models (7 nomograms, 1 equation-only), and 4 comparative machine learning studies testing multiple algorithms including ensemble methods. No study achieved low risk of bias globally using PROBAST+AI, with 83,3% classified as high risk and 16,7% with uncertain risk. Geographic concentration was extreme, with 91.7% of studies conducted in China, and no studies performed external validation. Although models demonstrated apparently satisfactory discriminative performance (AUC 0.72-0.988), 66.7% presented inadequate events-per-variable ratios (<10). The most frequent predictors were diabetes mellitus and BMI/obesity (58.3% each), age (41.7%), and serum albumin (41.7%).

Conclusions: The absence of methodologically robust studies, extreme geographic concentration, and lack of external validation prevent definitive conclusions about the clinical efficacy of predictive models for SSI in lumbar spine surgery. Clinical implementation would be premature without rigorous external validation in diversified populations. Future research should prioritize prospective multicenter studies with adequate samples ($EPV \geq 10$) and mandatory external validation.

Keywords: surgical site infection, lumbar spine surgery, predictive models, machine learning, nomograms, systematic review

Resumo

Contexto: As infecções do local cirúrgico após cirurgia de coluna lombar representam uma complicação significativa com impacto clínico e económico substancial. Os modelos preditivos emergiram como ferramentas potenciais para estratificação de risco, mas a sua qualidade metodológica e aplicabilidade clínica permanecem pouco claras.

Métodos: Foi conduzida uma revisão sistemática seguindo as diretrizes PRISMA com protocolo registado no PROSPERO (CRD42025613496). Pesquisámos *PubMed*, *SCOPUS*, *EBSCO*, *ScienceDirect* e *Cochrane Library* para estudos que desenvolveram ou validaram modelos preditivos para ILC em cirurgia de coluna lombar. A qualidade metodológica foi avaliada usando a ferramenta PROBAST+AI para avaliação do risco de viés de modelos preditivos. A extração de dados incluiu características dos estudos, métricas de performance dos modelos e variáveis preditoras.

Resultados: Doze estudos foram incluídos, compreendendo 8 modelos de regressão logística (7 nomogramas, 1 apenas equação) e 4 estudos comparativos de *machine learning* testando múltiplos algoritmos incluindo métodos *ensemble*. Nenhum estudo alcançou baixo risco de viés globalmente usando PROBAST+AI, com 83,3% classificados como alto risco e 16,7% com risco incerto. A concentração geográfica foi extrema, com 91,7% dos estudos conduzidos na China e nenhum estudo realizou validação externa. Embora os modelos demonstrassem performance discriminativa aparentemente satisfatória (AUC 0,72-0,988), 66,7% apresentaram rácios eventos-por-variável inadequados (<10). Os preditores mais frequentes foram diabetes *mellitus* e IMC/obesidade (58,3% cada), idade (41,7%) e albumina sérica (41,7%).

Conclusões: A ausência de estudos metodologicamente robustos, concentração geográfica extrema e falta de validação externa impedem conclusões definitivas sobre a eficácia clínica dos modelos preditivos para ILC em cirurgia de coluna lombar. A implementação clínica seria prematura sem validação externa rigorosa em populações diversificadas. Investigação futura deve priorizar estudos prospetivos multicêntricos com amostras adequadas ($EPV \geq 10$) e validação externa obrigatória.

Palavras-chave: infecção do local cirúrgico, cirurgia de coluna lombar, modelos preditivos, *machine learning*, nomogramas, revisão sistemática

Declaração

Resultados Científicos

No contexto desta dissertação, ainda não foram publicados resultados científicos. Contudo, encontra-se em preparação a submissão de um artigo científico derivado do trabalho desenvolvido.

Declaração Financeira

A presente dissertação foi realizada sem qualquer tipo de apoio financeiro.

Declaração de uso de IA generativa no processo de escrita

Reconhece-se que o autor, utilizou o apoio de ferramentas tecnológicas de inteligência artificial para revisão de texto e formatação, nomeadamente o GPT-4/5. Isto para garantir clareza e coerência no texto.

Índice

<i>Agradecimentos</i>	v
<i>Abstract</i>	vii
<i>Resumo</i>	ix
Declaração	xi
Índice	xiii
<i>Lista de Imagens</i>	xvi
<i>Lista de Tabelas</i>	xviii
<i>Acrónimos</i>	xx
1. Introdução	3
2. Objetivos	9
3. Motivação	12
4. Materiais e Métodos	16
4.1 Estratégia de pesquisa	16
4.2 Critérios de Elegibilidade	17
4.3 Seleção de estudos	18
4.4 Extração e análise de dados	18
4.5 Avaliação da Qualidade Metodológica dos Estudos	19
4.6 Considerações éticas	20
5. Resultados	23
5.1 Seleção de artigos	23
5.2 Avaliação da qualidade metodológica dos estudos com PROBAST+AI	25
5.3 Características dos estudos	26
5.4 Modelos Preditivos	31
5.4.1 Modelos Baseados em Regressão Logística	31
5.4.1.1 Nomogramas	31
5.4.1.2 Modelo com Equação Matemática	32
5.4.2 Modelos de Machine Learning	33
5.5 Variáveis preditivas	34
6. Discussão	38
6.1 Qualidade Metodológica e Risco de Viés	38
6.2 Concentração geográfica e limitações de generalização	39
6.3 Performance Preditiva	40
6.4 Variáveis Preditivas	42

6.5 Implicações para a Prática Clínica	42
6.6 Limitações desta Revisão	43
7. Conclusão	47
8. Recomendações e estudos futuros	49
9. Referências	52
10. Anexo	60
Anexo 1: Formulário De Seleção	60
Anexo 2: Formulário De Elegibilidade	62

Lista de Imagens

Figura 5.1: Fluxo de seleção dos estudos.....	24
Figura 5.2: Gráfico da Análise do Risco de Vies com PROBAST+AI.....	25

Lista de Tabelas

Tabela 4.1	Estratégia de pesquisa: queries de cada base de dados	17
Tabela 5.1	Motivos de exclusão após leitura completa.	24
Tabela 5.2	Resultado da aplicação do PROBAST+AI (Domínios- D1: Participantes, D2:Preditores, D3: resultados, D4: Análise)	26
Tabela 5.3	Características dos estudos que desenvolveram nomogramas baseados em regressão logística	28
Tabela 5.4	Características do estudo baseado em regressão logística, com equação matemática. . .	29
Tabela 5.5	Características dos estudos de Machine Learning	30

Acrónimos

ASA American Society of Anesthesiologists

AUC Area Under the Curve

CDC Centers for Disease Control and Prevention

CV Cross-Validation

DCA Decision Curve Analysis

DCC Department of Computer Science

DT Decision Tree

EPV Eventos Por Variável

FCUP Faculty of Sciences of the University of Porto

FMUP Faculty of Medicine of the University of Porto

FPG-CV Coeficiente de Variação da Glicemia em Jejum

GBM Gradient Boosting Machine

HbA_{1c} Hemoglobina Glicada

IA Inteligência Artificial

ILC Infecções do Local Cirúrgico

IMC Índice de Massa Corporal

KNN K-Nearest Neighbors

LASSO Least Absolute Shrinkage and Selection Operator

LCR Líquido Cefalorraquidiano

LR Logistic Regression

MEDCIDS Department of Health Information and Decision Sciences

MeSH Medical Subject Headings

MIS-LIF Minimally Invasive Lumbar Interbody Fusion

ML Machine Learning

MLP Multilayer Perceptron

NB Naive Bayes

OR Odds Ratio

PICO Population, Intervention, Comparison, Outcome

PLIF Posterior Lumbar Interbody Fusion

PRISMA Preferred Reporting Items for Systematic Reviews and Meta-Analyses

PROBAST+AI Prediction Model Risk of Bias Assessment Tool + Artificial Intelligence

PROSPERO International Prospective Register of Systematic Reviews

RF Random Forest

RR Risco Relativo

SMOTE - Synthetic Minority Oversampling Technique

SVM-RFE - Support Vector Machine - Recursive Feature Elimination

SVM Support Vector Machine

TLIF Transforaminal Lumbar Interbody Fusion

XGB Extreme Gradient Boosting



Introdução

1. Introdução

A doença degenerativa da coluna vertebral constitui uma das patologias crónicas mais prevalentes em populações envelhecidas. A mesma compromete significativamente a qualidade de vida através da dor crónica e incapacidade física. Embora muitos casos possam ser tratados conservadoramente, observa-se uma crescente necessidade de intervenção cirúrgica com instrumentação espinal (Park et al., 2022).

Como resultado do progressivo envelhecimento da população, prevê-se então que mais pacientes necessitam de intervenção cirúrgica, particularmente de fusão lombar (Shah et al., 2021; Park et al., 2022). A dimensão desta problemática é evidenciada pelo facto de se realizarem mais de 450.000 cirurgias dessas, anualmente apenas nos Estados Unidos (Shah et al., 2021).

As infeções do local cirúrgico (ILC) são definidas como infeções que ocorrem no local da incisão e/ou nos tecidos e órgãos subjacentes mais profundos, manifestando-se dentro de 30 dias após o procedimento cirúrgico ou até 90 dias em cirurgias que envolvem a colocação de próteses (Borchardt & Tzizik, 2018). De acordo com dados da Direção-Geral da Saúde (2022), as ILC estão entre as infeções associadas aos cuidados de saúde mais prevalentes no contexto português. As ILC após cirurgia da coluna lombar, podem resultar em falha de fixação, osteomielite (Janssen, et al., 2019), lesão neurológica e pseudo-artrose (Mueller et al., 2024). Mas as consequências não se ficam por aí, para além destas sequelas físicas diretas, que comprometem significativamente o prognóstico dos doentes. As ILC desencadeiam uma cascata de complicações: internamentos prolongados, necessidade de intervenções cirúrgicas adicionais, eventual transferência para unidades de cuidados intensivos, bem como ao aumento das taxas de mortalidade e morbilidade (DGS, 2022). Este conjunto de complicações gera ainda custos de saúde extraordinariamente elevados, impondo também um pesado encargo económico para os pacientes, as suas famílias e a sociedade em geral (Matsumoto et al., 2022).

A incidência de ILC varia consideravelmente, dependendo das características individuais do paciente e da complexidade do procedimento cirúrgico (Lubelski et al., 2021). Especificamente, no que se refere às infeções do local cirúrgico em cirurgias da coluna vertebral, a incidência pode variar entre 0,7% e 16%, sendo que as infeções na região lombar representam 51% do total de casos (Ge et al., 2024).

Esta diferença na incidência é explicada por múltiplos fatores determinantes bem documentados na literatura científica. A complexidade do procedimento é determinante, com descompressão isolada a apresentar taxas significativamente menores em técnicas minimamente invasivas comparativamente à cirurgia aberta (0,4% vs. 3,9%), e mesmo os procedimentos de descompressão com fusão mantêm diferenças substanciais de taxa ILC com 0,7%, para técnicas minimamente invasivas e 2,6% em cirurgia aberta (Mueller et al., 2019).

O diagnóstico precoce é desafiante, uma vez que, frequentemente, não existem sinais ou sintomas específicos que indiquem claramente a presença de infecção (Janssen, et al., 2019). Quando presentes, os sintomas podem incluir dor, febre, sensibilidade e rubor local (Pei et al., 2021). Por sua vez, o tratamento revela-se igualmente complexo, exigindo frequentemente múltiplos desbridamentos cirúrgicos combinados com terapêutica antibiótica prolongada (Janssen, et al., 2019). Mais preocupante ainda é a crescente prevalência de microrganismos resistentes aos antibióticos, o que torna o tratamento destas infecções progressivamente mais desafiante (Janssen, et al., 2019). Sublinha-se assim a importância da sua prevenção.

Neste contexto, diversas investigações têm revelado que o risco de ILC após cirurgia da coluna vertebral é influenciado por múltiplos fatores de risco. Entre os mais consistentemente identificados destacam-se, a nível pré-operatório: a idade avançada, histórico prévio de cirurgia da coluna, obesidade, diabetes mellitus, tabagismo, hipertensão e estado nutricional deficiente (Janssen, et al., 2019; Lubelski et al., 2021). A nível intraoperatório a perda sanguínea significativa, o número de níveis vertebrais instrumentados e duração prolongada do procedimento cirúrgico (Janssen, et al., 2019; Lubelski et al., 2021). Curiosamente, estudos recentes têm também identificado fatores menos óbvios como a angina de peito, cancro metastático, infecções musculoesqueléticas prévias e histórico de hemorragia cerebral (Shah et al., 2021).

Contudo, tradicionalmente, estes fatores de risco são apresentados como riscos relativos (RR) ou odds ratios (OR). No entanto, esta abordagem não é suficiente para calcular o risco individualizado de cada doente quando se considera a interação complexa entre múltiplos fatores (Janssen, et al., 2019). Na prática clínica, os cirurgiões experientes conseguem intuir se um paciente apresenta maior ou menor risco de infecção. No entanto, esta estimativa torna-se progressivamente mais imprecisa e sujeita a vieses quando o doente apresenta características contraditórias. Como por exemplo: como avaliar um não-fumador obeso versus um fumador com peso normal (Lubelski et al., 2021).

Para responder a esta necessidade de previsão individualizada de risco, os nomogramas têm-se estabelecido como ferramentas práticas na medicina preditiva. Os nomogramas são representações gráficas de fórmulas matemáticas complexas, amplamente utilizados como dispositivos prognósticos em oncologia e medicina (Balachandran et al., 2015). Estes instrumentos utilizam variáveis biológicas e clínicas, para gerar um modelo estatístico prognóstico que calcula uma probabilidade individualizada de um evento clínico específico, tal como recidiva de cancro (Balachandran et al., 2015). A sua principal vantagem reside na capacidade de reduzir modelos preditivos estatísticos complexos numa única estimativa numérica da probabilidade de um evento, apresentada através de interfaces gráficas amigáveis que facilitam a utilização durante consultas clínicas para gerar uma tomada de decisão clínica informada (Iasonos et al., 2008). Mais recentemente, tem-se observado uma crescente integração entre nomogramas tradicionais e técnicas avançadas de *machine learning* (ML). Onde algoritmos como *Least Absolute Shrinkage and Selection Operator* (LASSO) são utilizados para seleção de variáveis preditivas, permitindo comparações diretas entre o desempenho de nomogramas convencionais (baseados na regressão logística simples) e modelos de *machine learning* mais sofisticados na mesma população clínica (Gu & Lu, 2025).

Estes algoritmos de *machine learning*, enquanto ramo da inteligência artificial, oferecem capacidades preditivas e exploratórias avançadas. Estes métodos utilizam algoritmos sofisticados que simulam o processo de aprendizagem humana, melhorando continuamente a sua precisão através da análise de grandes volumes de dados (Kim et al., 2018). Esta tecnologia tem encontrado aplicação crescente no campo da medicina, demonstrando frequentemente precisão preditiva superior aos métodos convencionais (Kim et al., 2018).

Surpreendentemente, verifica-se que poucos trabalhos de investigação têm explorado efetivamente a integração dos fatores de risco conhecidos com algoritmos de ML para identificar pacientes com elevado risco de desenvolver ILC, particularmente no contexto específico da cirurgia de coluna lombar (Hopkins et al., 2020).

Apesar destes avanços promissores, existem limitações importantes na literatura atual que comprometem a implementação clínica efetiva destes modelos. Um dos principais obstáculos é a heterogeneidade metodológica, com os estudos existentes a utilizarem definições inconsistentes de ILC e critérios de inclusão díspares (Lauinger et al., 2024). Esta heterogeneidade dificulta a comparação entre a literatura e a generalização dos resultados.

Além disso, muitos artigos baseiam-se em bases de dados administrativas que dependem de códigos de diagnóstico, o que pode subestimar as taxas de complicação e carecer da granularidade necessária para uma avaliação precisa (Shah et al., 2021). Os próprios autores reconhecem a existência destas limitações no uso de bases de dados retrospectivas e na dependência de códigos de diagnóstico (Shah et al., 2021). Além disso, a validação externa está frequentemente ausente, elevando o risco de *overfitting* e reduzindo a capacidade de generalizar para novas coortes, limitando assim a sua aplicabilidade clínica (Shah et al., 2021). Esta limitação é particularmente relevante dado o risco de *overfitting*, onde o algoritmo apresenta boa performance no conjunto de desenvolvimento, mas generalização diminuída para novas coortes (Shah et al., 2021).

Acresce ainda que, as investigações são frequentemente limitadas a instituições específicas ou regiões geográficas. Como demonstrado pelo trabalho de Shah et al. (2021) que se baseou exclusivamente em dados da Califórnia, não considerando as diferenças nos sistemas de saúde, práticas cirúrgicas e características populacionais que podem afetar o desempenho dos modelos. Adicionalmente, Peng et al. (2019) reconheceram que o tamanho limitado da amostra de casos com infeção comprometeu a capacidade de identificação completa dos fatores de risco. O que poderia ter levado a uma análise inadequada de alguns fatores e à não identificação de outros fatores de risco relevantes. Os mesmos autores recomendam então trabalhos de investigação com amostras de maior dimensão para validação dos resultados (Peng et al., 2019). Por último, embora alguns modelos de risco pré-operatório tenham sido validados e até implementados clinicamente com sucesso, a maioria dos modelos propostos não completou validação externa adequada, e futuras tentativas de desenvolvimento devem considerar aspetos práticos de implementação, incluindo usabilidade e integração com sistemas clínicos existentes (Lauinger et al., 2024).

Estas limitações metodológicas e de validação evidenciam uma lacuna crítica na literatura atual, sublinhando a necessidade de desenvolver modelos preditivos robustos capazes de avaliar o risco individualizado, proporcionando assim uma base sólida para o aconselhamento pré-operatório dos doentes. Afinal, prevenir e gerir complicações pós-operatórias é fundamental na prática clínica atual, dado que o seu impacto para o doente é substancial: maior risco de mortalidade, morbilidade prolongada e custos hospitalares elevados (Lim, et al., 2018). Este ponto é particularmente relevante em cirurgias eletivas, onde uma avaliação precisa do risco pode influenciar a técnica, a preparação do doente ou a própria decisão entre tratamento conservador e intervenção cirúrgica (Lubelski et al., 2021).

Apesar dos avanços significativos na aplicação de modelos preditivos em diversas áreas da medicina, persiste uma escassez no conhecimento relativamente à qualidade metodológica e performance destes modelos na previsão de ILC em cirurgias de coluna lombar. Torna-se, por isso, necessária uma avaliação sistemática e rigorosa da evidência científica disponível, para caracterizar o desempenho, identificar as

limitações existentes e avaliar a aplicabilidade clínica real destes modelos, fornecendo orientações para a sua implementação responsável na prática clínica. Face a esta lacuna identificada na evidência científica, a presente revisão sistemática procura responder à seguinte questão de investigação:

"Qual o impacto dos modelos preditivos desenvolvidos para estratificação do risco de infeções do local cirúrgico em pacientes submetidos a cirurgia de coluna lombar, considerando a sua performance preditiva e potencial para implementação clínica?"

Esta questão, estruturada segundo o formato PICO (População: pacientes submetidos a cirurgia de coluna lombar; Intervenção: modelos preditivos para ILC; Comparação: diferentes tipos de modelos preditivos e suas abordagens metodológicas; *Outcome*: impacto na previsão de ILC, performance preditiva e viabilidade de implementação clínica).

A dissertação encontra-se estruturada em oito capítulos principais. Após esta introdução, são apresentados os objetivos específicos desta investigação. Já no capítulo seguinte encontra-se a Motivação, onde é apresentada a relevância e o porquê do desenvolvimento deste trabalho. O Capítulo 4, “Materiais e Métodos” descreve detalhadamente a metodologia da revisão sistemática, incluindo a estratégia de pesquisa, critérios de elegibilidade, processo de seleção de artigos e métodos de avaliação da qualidade (com o PROBAST+AI). De seguida, apresentam-se os resultados obtidos através da análise sistemática da literatura, caracterizando os estudos incluídos e resumindo os resultados de desempenho dos modelos preditivos. Posteriormente, a “Discussão”, onde se analisa criticamente os resultados no contexto da literatura existente, explorando as implicações clínicas, limitações identificadas e direções futuras de investigação. Segue-se o capítulo “Conclusão” onde são conclusões principais e as recomendações para a prática clínica e investigação futura. Por fim, no último capítulo são apresentadas as recomendações para estudos futuros.



Objetivos

2. Objetivos

O objetivo principal deste estudo é: avaliar a qualidade metodológica e performance preditiva dos modelos desenvolvidos para estratificação do risco de infecções do local cirúrgico em pacientes submetidos a cirurgia de coluna lombar.

Tendo, no entanto, como objetivos secundários:

- Identificar, através de uma revisão sistemática da literatura, artigos que desenvolveram ou validaram modelos preditivos na avaliação do risco de infecção do local cirúrgico nas cirurgias de coluna lombar.
- Caracterizar os modelos preditivos identificados, considerando a sua metodologia, o tipo de modelo, variáveis utilizadas e métricas de desempenho.
- Avaliar a qualidade metodológica dos estudos incluídos utilizando a ferramenta PROBAST+AI (*Prediction model Risk Of Bias Assessment Tool*).
- Analisar e sintetizar as métricas de performance reportadas, como AUC, sensibilidade, especificidade, calibração e outras métricas, se disponíveis.
- Avaliar a aplicabilidade clínica e limitações dos modelos preditivos na prática cirúrgica.



Motivação

3. Motivação

Apesar dos avanços significativos no conhecimento dos fatores de risco para infecções do local cirúrgico em cirurgias de coluna lombar, persiste uma lacuna crítica entre o conhecimento teórico e a sua aplicação prática na avaliação individualizada do risco.

Como já referido, a abordagem tradicional, baseada em *odds* ratios e riscos relativos isolados, revela-se insuficiente quando os clínicos enfrentam pacientes com perfis de risco contraditórios (Janssen, et al., 2019). Esta limitação tem implicações diretas na qualidade dos cuidados e nos *outcomes* dos pacientes, particularmente num contexto de crescimento exponencial das cirurgias de fusão lombar (Park et al., 2022).

As ILC da coluna vertebral apresentam consequências particularmente severas e resultam num aumento de custos hospitalares, decorrente de internamentos prolongados e da necessidade de múltiplas re-intervenções cirúrgicas (Lupo et al., 2025). Estes custos podem variar entre USD \$15.817 e \$38.701 por caso (Lupo et al., 2025), representando aumentos de até 65% nos custos hospitalares (Atkinson et al., 2017).

Adicionalmente, as transformações recentes no setor da saúde, designadamente a adoção e implementação do conceito de *Value-Based Healthcare*, têm desafiado os prestadores de cuidados de saúde a alcançar simultaneamente melhor qualidade assistencial e redução de custos por doente tratado (Matsumoto et al., 2022). Neste contexto, as ILC assumem particular relevância por se caracterizarem como complicações simultaneamente frequentes, dispendiosas e potencialmente evitáveis.

Assim sendo, a identificação precoce de doentes em alto risco antes da intervenção cirúrgica torna-se essencial para permitir um planeamento cirúrgico mais eficaz e personalizado (Matsumoto et al., 2022). A quantificação precisa do risco individual pode auxiliar as instituições hospitalares a estratificar mais eficientemente os seus doentes. Isto promove, tanto comparações mais significativas entre resultados institucionais, como a distribuição mais equitativa dos recursos, baseada nas necessidades específicas das populações (Matsumoto et al., 2022).

Apesar do sucesso dos modelos preditivos noutras áreas médicas, a sua aplicação na previsão de ILC em cirurgias de coluna lombar permanece limitada. Ademais, a literatura existente apresenta heterogeneidade metodológica, dependência de bases de dados administrativas e ausência de validação externa, comprometendo a implementação clínica.

É importante ainda referir que, embora existam revisões sistemáticas sobre modelos preditivos para ILC em cirurgias da coluna vertebral em geral (Lauinger et al., 2024), não existe até à data nenhuma revisão sistemática especificamente focada na coluna lombar. Esta lacuna é particularmente relevante, considerando

que as cirurgias lombares apresentam características anatómicas, técnicas cirúrgicas e perfis de risco distintos das restantes regiões da coluna vertebral. Além disso, estudos demonstram taxas de infecção significativamente diferentes entre regiões da coluna, com a região lombar (e toracolombar) a apresentar maior incidência de ILC comparativamente às regiões cervical e torácica (Alfin et al., 2025).

Face a esta necessidade identificada, torna-se imperativo realizar uma avaliação sistemática da evidência disponível sobre o impacto dos modelos preditivos na avaliação do risco de ILC em cirurgias de coluna lombar. Os resultados desta investigação beneficiarão diretamente cirurgiões na tomada de decisão clínica, administradores hospitalares na otimização de recursos, e fundamentalmente, os pacientes através de cuidados mais seguros e personalizados.



Materiais e Métodos

4. Materiais e Métodos

Esta revisão sistemática foi desenvolvida de acordo com as diretrizes estabelecidas pelo *Preferred Reporting Items for Systematic Reviews and Meta-Analyses* (PRISMA) (Page et al., 2021). A estratégia de pesquisa bibliográfica foi realizada nas seguintes bases de dados: *PubMed*, *SCOPUS*, *EBSCO*, *SCIENCEDIRECT* e *Cochrane Library*. Relativamente ao protocolo da revisão sistemática, o mesmo foi submetido e publicado na base de dados PROSPERO, a 17 Janeiro de 2025, sob o número de registo CRD42025613496.

4.1 Estratégia de pesquisa

A pesquisa de artigos foi realizada entre 1 e 4 de Fevereiro de 2025 nas seguintes bases de dados eletrónicas: *PubMed*, *SCOPUS*, *EBSCO*, *ScienceDirect* e *Cochrane Library*. Foram consultadas também as bases de dados: IEEE Xplore e Web of Science, no entanto estas últimas não apresentaram artigos que atendessem a *query* estabelecida.

A pesquisa foi conduzida sem restrições temporais, incluindo todos os estudos disponíveis até à data da pesquisa (fevereiro de 2025), de forma a capturar a evolução histórica dos modelos preditivos na área.

As *queries* para a realização da pesquisa foram desenvolvidas com base na pergunta PICO e incluíram uma combinação de termos *MeSH* (*Medical Subject Headings*) e palavras-chave em texto livre, como: "*surgical site infection*", "*lumbar spine surgery*", "*predictive models*" e "*machine learning*". Foi aplicado o operador booleano "AND" para combinar os descritores de conceitos diferentes, enquanto o operador "OR" foi utilizado para incluir sinónimos dentro de cada conceito.

Por forma a garantir maior abrangência dos estudos, foram utilizadas ambas as grafias para o mesmo procedimento cirúrgico: "*discectomy*" (tradicionalmente americana) e "*discectomy*" (britânica) (Dorow et al., 2017).

As estratégias de pesquisa completas para todas as bases de dados consultadas encontram-se detalhadas na tabela 4.1, tendo sido adaptadas às especificidades de cada plataforma mantendo o significado dos conceitos centrais.

Tabela 4.1 Estratégia de pesquisa: queries de cada base de dados

PUBMED	(((((machine learning) OR (statistical model*)) OR (predictive model*)) AND (((surgical site infection) OR (surgical wound infection)) OR (cross infection))) AND (((((lumbar spine) OR (spinal fusion)) OR (laminectomy)) OR (discectomy)) OR (discectomy)))
SCOPUS	(TITLE-ABS-KEY ((machine AND learning) OR (statistical AND model) OR (predictive AND model)) AND ALL ((surgical AND (site OR wound) AND infection) OR (cross AND infection)) AND TITLE-ABS-KEY ((lumbar AND spine) OR (spinal AND fusion) OR (laminectomy) OR (discectomy) OR (discectomy))))
SCIENCEDIRECT	("lumbar spine" OR "spinal fusion" OR "laminectomy" OR "discectomy") AND ("surgical wound infection" OR "surgical site infection") AND ("machine learning" OR "statistical model" OR "predictive model") ("lumbar spine" OR "spinal fusion" OR "laminectomy" OR "discectomy") AND ("surgical wound infection" OR "surgical site infection") AND ("machine learning" OR "statistical model" OR "predictive model")
EBSCO	("machine learning" OR "artificial intelligence" OR "predictive model" OR "statistical model") AND ("surgical site infection" OR "surgical wound infection" OR "cross infection" OR "postoperative infection") AND ("lumbar spine" OR "spinal fusion" OR "laminectomy" OR "discectomy" OR "discectomy" OR "spinal surgery")
Cochrane	(((machine AND learning) OR (statistical AND model) OR (predictive AND model)) AND ((surgical AND (site OR wound) AND infection) OR (cross AND infection)) AND ((lumbar AND spine) OR (spinal AND fusion) OR (laminectomy) OR (discectomy) OR (discectomy))))

4.2 Critérios de Elegibilidade

No que diz respeito a elegibilidade dos estudos para inclusão nesta revisão sistemática, os mesmos atenderam aos seguintes critérios.

Relativamente à população, foram eleitos estudos em pacientes adultos com idade igual ou superior a 18 anos, submetidos a cirurgia de coluna lombar eletiva. Tendo sido considerados os procedimentos cirúrgicos para condições degenerativas, traumáticas ou correção de deformidades da coluna vertebral.

Foram ainda incluídos estudos que avaliam modelos preditivos, algoritmos ou ferramentas de avaliação de risco especificamente desenvolvidos ou validados para avaliar o risco de ILC em pacientes submetidos a cirurgia de coluna lombar.

Relativamente ao desenho de estudo, foram considerados ensaios clínicos randomizados, estudos de coorte prospectivos ou retrospectivos, estudos transversais e investigação qualitativa que forneçam dados empíricos sobre a eficácia de modelos preditivos. Adicionalmente, foram incluídos apenas artigos originais publicados em revistas científicas com revisão por pares, redigidos em inglês ou português. Não existiu restrição temporal de publicação, tendo sido incluídos todos os estudos disponíveis até à data da pesquisa.

Foram excluídos da revisão os estudos que apresentavam pacientes com infeções pré-existentes no momento da cirurgia ou pacientes submetidos a cirurgias de emergência ou trauma agudo que requeressem intervenção imediata. Foram também eliminados estudos que não avaliem especificamente modelos preditivos para ILC, ou que se foquem exclusivamente em intervenções terapêuticas para tratamento de infeções já adquiridas.

Quanto ao tipo de publicação, foram excluídos artigos de opinião, editoriais, cartas ao editor, resumos de conferências, capítulos de livros, teses, dissertações e literatura cinzenta. Foram ainda eliminados estudos sem dados empíricos quantitativos ou qualitativos relevantes sobre o desempenho de modelos preditivos, bem como estudos que abordem modelos preditivos para outras complicações pós-operatórias que não ILC, ou modelos sem análise específica para cirurgia de coluna lombar.

Para garantir a consistência na seleção e análise dos estudos, foram estabelecidas as seguintes definições: as infecções do local cirúrgico foram consideradas segundo os critérios de diagnóstico estabelecidos pelo Centers for Disease Control and Prevention (CDC), com especificação temporal de ocorrência entre 30 a 90 dias pós-operatório, conforme referido acima. A cirurgia de coluna lombar foi definida como procedimentos cirúrgicos realizados na região anatômica L1-L5, incluindo fusão lombar (TLIF, PLIF, MIS-LIF), laminectomia, discectomia, e outros procedimentos corretivos da coluna lombar. Os modelos preditivos foram caracterizados como ferramentas estatísticas ou algoritmos com capacidade de prognóstico individualizado, capazes de fornecer estimativas de risco específicas para cada paciente com base em variáveis preditoras definidas (Iwagami & Matsui, 2022; van Smeden, 2021).

4.3 Seleção de estudos

A seleção dos estudos foi realizada em duas fases por dois revisores independentes, utilizando formulários específicos desenvolvidos para cada uma das fases, presente no anexo 1 e 2, respetivamente. Na primeira fase, análise de títulos e resumos, foi aplicado um formulário de seleção mais abrangente que incluía critérios gerais como a avaliação de cirurgias de coluna vertebral (não restringindo especificamente à região lombar), a presença de modelos preditivos e a população adulta. Esta abordagem, mais inclusiva, pretendia evitar a exclusão prematura de estudos cujos títulos ou resumos não explicitassem claramente o foco na coluna lombar. Esses artigos poderiam conter dados relevantes se, por exemplo, abordassem múltiplas regiões da coluna vertebral e apresentassem a análise de dados separadas por segmento anatômico (lombar, cervical, torácico).

Na segunda fase, para avaliação do texto completo dos artigos, foi utilizado um formulário de elegibilidade mais específico e restrito. Este formulário afunilava os critérios para cirurgias especificamente da coluna lombar e modelos preditivos com métricas de desempenho claramente definidas.

Esta estratégia permitiu a realização de uma triagem inicial mais ampla, seguida de uma avaliação mais rigorosa e específica. Tentamos assim, garantir que nenhum estudo potencialmente relevante fosse inadvertidamente excluído durante o processo de seleção inicial.

O processo de seleção foi documentado utilizando o diagrama PRISMA (Page et al., 2021), assegurando transparência e rastreabilidade.

4.4 Extração e análise de dados

Desenvolvemos, especificamente para esta revisão sistemática, um formulário padronizado para a extração de dados. O formulário foi estruturado para recolher informações abrangentes sobre as

características dos estudos incluídos. As mesmas foram organizadas em categorias específicas que permitissem uma análise sistemática e comparativa dos modelos preditivos encontrados.

As informações extraídas incluíram dados bibliográficos dos estudos, nomeadamente autores, país de origem e ano de publicação, bem como o título completo dos artigos. Foi registado também o tipo de estudo, a população estudada incluindo o tamanho da amostra, características demográficas dos participantes (percentagem de participantes do sexo masculino e feminino). No que se refere aos modelos preditivos, foram extraídos dados sobre o tipo de modelo/intervenção utilizada, as variáveis incluídas nos modelos preditivos, os métodos de desenvolvimento e validação empregados, e por fim os desfechos principais avaliados. As métricas de desempenho reportadas pelos estudos foram extraídas conforme disponibilidade, incluindo principalmente: área sob a curva ROC (AUC), sensibilidade, especificidade, e quando reportadas pelos autores, outras métricas como acurácia e medidas de calibração.

Devido a diversidade metodológica dos estudos incluídos e os diferentes tipos de modelos preditivos avaliados, foi realizada uma síntese narrativa dos resultados. Esta escolha possibilitou uma análise qualitativa mais abrangente, comparando as características dos modelos, variáveis preditivas, indicadores de performance e limitações reportadas.

A análise de dados foi executada de forma a identificar padrões comuns entre os estudos, variações nas abordagens metodológicas utilizadas para o desenvolvimento de modelos preditivos, e diferenças nas métricas de desempenho reportadas. Os resultados foram organizados e apresentados em tabelas descritivas que facilitam a comparação entre os diferentes modelos e estudos incluídos. Adicionalmente, foi realizada uma análise das variáveis preditivas mais frequentemente identificadas como significativas nos diferentes modelos, bem como uma avaliação das limitações metodológicas mais comuns reportadas pelos autores.

A síntese narrativa foi escolhida devido à heterogeneidade metodológica significativa entre os estudos: diversidade nos tipos de modelos, variabilidade nas métricas reportadas, diferenças nas metodologias de desenvolvimento. Esta heterogeneidade impossibilitou a meta-análise quantitativa, justificando a síntese narrativa estruturada.

4.5 Avaliação da Qualidade Metodológica dos Estudos

Para avaliar a qualidade metodológica dos estudos incluídos, optou-se pela ferramenta PROBAST+AI (Prediction Model Risk of Bias Assessment Tool + Artificial Intelligence), uma vez que esta é a versão atualizada do PROBAST original (Moons et al., 2025). A mesma foi aplicada independentemente por dois avaliadores em todos os estudos. As divergências de avaliação por domínio foram resolvidas por consenso entre os dois.

Enquanto o PROBAST original foi desenvolvido primariamente para avaliar modelos de regressão estatística tradicional, o PROBAST+AI incorpora critérios específicos para a avaliação de algoritmos de *machine learning e inteligência artificial* (Moons et al., 2025). Esta ferramenta inclui questões sinalizadoras adicionais que nos permitem abordar aspetos metodológicos críticos dos modelos de IA, como a adequação do tamanho amostral, estratégias de prevenção de *overfitting* e métodos de validação apropriados para modelos não-lineares (Moons et al., 2025). Embora mantenha a estrutura fundamental de quatro domínios do

PROBAST original (participantes, preditores, outcome e análise), o PROBAST+AI permite uma avaliação mais abrangente e metodologicamente rigorosa da qualidade dos estudos (Moons et al., 2025). Isto aplica-se tanto a abordagens estatísticas convencionais como a algoritmos de *machine learning*, possibilitando assim uma análise mais precisa da evidência disponível sobre modelos preditivos para ILC em cirurgia de coluna lombar.

A avaliação foi realizada através da análise dos quatro domínios principais, sendo que os três primeiros domínios foram também avaliados quanto à aplicabilidade dos resultados à questão de investigação desta revisão sistemática. Cada domínio foi classificado como tendo preocupação baixa, alta ou incerta com base nas questões sinalizadoras. A classificação geral foi determinada seguindo critérios pré-definidos: baixa quando todos os domínios apresentaram preocupação baixa, alta quando pelo menos um domínio apresentou preocupação alta, e incerta quando pelo menos um domínio foi classificado como incerto e nenhum como alto (Moons et al., 2025).

Os resultados da avaliação da qualidade metodológica foram utilizados para informar a interpretação dos resultados da síntese narrativa, permitindo identificar potenciais limitações que poderiam afetar a confiabilidade das conclusões sobre a eficácia dos modelos preditivos avaliados.

4.6 Considerações éticas

Não foi necessária aprovação por comité de ética para a condução desta investigação, uma vez que se baseou exclusivamente na análise de dados já publicados na literatura científica. No entanto, todos os estudos incluídos na revisão foram previamente aprovados pelos respetivos comités de ética das instituições onde foram realizados, conforme documentado nas publicações originais.

Durante todo o processo de investigação, foram seguidos os princípios éticos de integridade científica. Tendo, o protocolo desta revisão sistemática sido previamente registado na base de dados PROSPERO antes do início da recolha de dados, garantindo transparência metodológica e reduzindo o risco de viés.



Resultados

5. Resultados

5.1 Seleção de artigos

Como referido anteriormente, a pesquisa de artigos nas bases de dados foi realizada entre 1 e 4 de Fevereiro de 2025. A mesma resultou na identificação de 1060 artigos. Destes, 125 foram encontrados na *PubMed*, 359 na SCOPUS, 18 na EBSCO, 434 na *ScienceDirect* e 124 na *Cochrane Library*. Não foram identificados artigos relevantes no *IEEE Xplore* e *Web of Science*, apesar da pesquisa ter sido realizada nestas bases de dados conforme a metodologia pré-estabelecida.

Na base de dados *ScienceDirect*, foi necessário realizar duas pesquisas separadas para garantir mais abrangência. Esta opção decorreu de uma limitação técnica da plataforma, que permite no máximo oito operadores booleanos por campo de pesquisa. Tendo isto em conta, e uma vez que "*discectomy*" e "*discectomy*" se referem ao mesmo procedimento cirúrgico, optou-se por separar ambos os termos em duas queries distintas. A primeira query incluiu então o termo "*discectomy*" e resultou em 208 artigos, enquanto a segunda incluiu "*discectomy*" e identificou 226 artigos. No total foram obtidos os 434 artigos nesta base de dados.

Seguidamente procedeu-se à exclusão de artigos duplicados no *Rayyan* (Ouzzani et al., 2016). Após a exclusão de 309 artigos duplicados, 751 referências foram analisadas com aplicação dos critérios de inclusão e exclusão na fase de triagem por título e resumo. Para auxiliar esta análise, foi aplicado o formulário de seleção, como explicado no capítulo anterior. Desta análise, 710 artigos foram excluídos por diferentes motivos que se enquadravam nos critérios de exclusão desta revisão sistemática, nomeadamente: estudos que apresentavam comparação entre uso de antibióticos; estudos aplicados à coluna completa e não especificamente à coluna lombar; estudos sem modelo preditivo (ex.: apenas faziam levantamento dos potenciais fatores de risco); uma revisão sistemática (que abordava modelos preditivos para avaliação de risco de infeção do local cirúrgico após cirurgia à coluna vertebral – não exclusivamente lombar); estudos envolvendo reinternamentos por infeção da ferida cirúrgica; e estudos com população pediátrica.

Os 41 artigos restantes foram considerados elegíveis para leitura completa e aplicação do formulário de elegibilidade. Durante esta fase, 29 artigos foram excluídos pelos motivos descritos na Tabela 5.1. Entre os artigos analisados, 5 geraram discordância inicial entre os revisores, sendo necessária discussão para consenso. Destes, 3 foram excluídos, enquanto 2 foram incluídos após discussão.

Assim, 12 artigos foram incluídos na revisão sistemática, como demonstrado na Figura 5.1 do fluxo de seleção dos estudos segundo as diretrizes PRISMA (Page et al., 2021).

Tabela 5.1: Motivos de exclusão após leitura completa

Motivo de Exclusão	n
Texto completo não disponível	2
Artigo não escrito em português ou inglês	1
Complicações pós-operatórias gerais (não ILC)	5
Coluna completa com complicações pós-operatórias gerais	7
Coluna completa (não exclusivamente lombar)	6
Revisão sistemática (sobre coluna completa)	1
Cirurgia toracolombar com complicações pós-operatórias gerais	2
Não apresentavam modelo preditivo	2
ILC em cirurgia toracolombar	3
Total de estudos excluídos	29

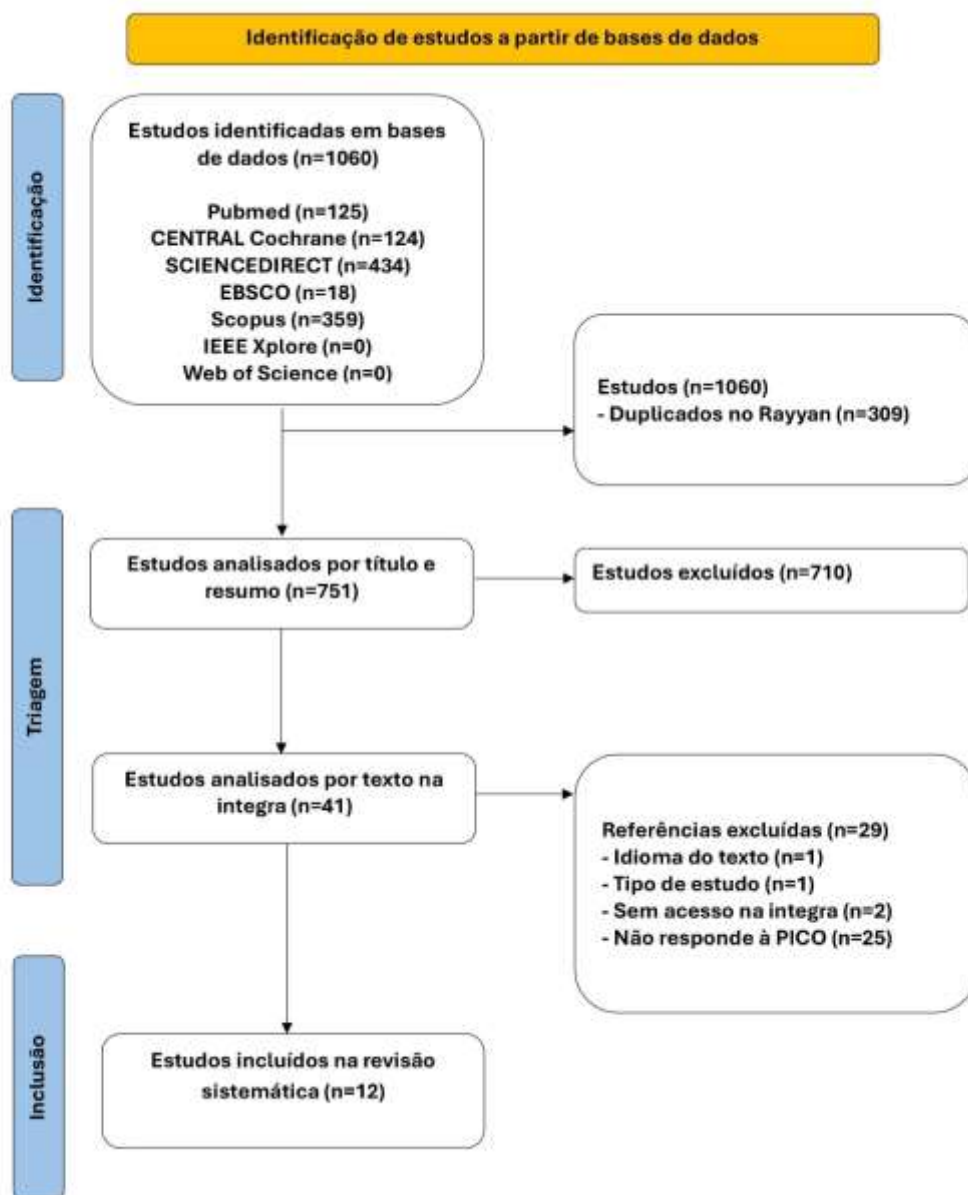


Figura 5.1: Fluxo de seleção dos estudos

5.2 Avaliação da qualidade metodológica dos estudos com PROBAST+AI

A avaliação da qualidade metodológica dos 12 estudos incluídos foi realizada aplicando a ferramenta PROBAST+AI, que avalia quatro domínios: participantes (D1), preditores (D2), resultados (D3) e análise (D4). Os resultados desta avaliação estão apresentados na Tabela 5.2 e Figura 5.2.

No primeiro domínio, oito estudos (66,7 %) apresentou baixa preocupação quanto à seleção e características dos participantes. Apenas quatro estudos (33,3%) demonstraram alto risco neste domínio, devido ao facto de os critérios de inclusão e exclusão não resultarem numa amostra representativa. Já no segundo e terceiro domínio, dos preditores e resultados, respetivamente, todos os estudos 12 estudos (100%) foram classificados como tendo risco incerto. A classificação como risco incerto nos domínios D2 e D3 deveu-se à ausência de informação reportada sobre o cegamento entre a avaliação dos preditores e dos resultados, impedindo uma avaliação conclusiva do risco de viés nestes domínios. Relativamente a análise (quarto domínio), este apresentou uma maior variabilidade, com 10 estudos (83,3%) classificados como alto risco de viés e apenas dois (16,7%) como risco incerto. O alto risco de viés neste domínio deveu-se principalmente ao tamanho amostral inadequado, com estudos a apresentar um rácio de eventos por variável (EPV) inferior a 10.

Posto isto, na avaliação geral do risco de viés, foram classificados 10 estudos (83,3%) como alto risco e apenas 2 estudos (16,7%) como risco incerto. É importante salientar, que nenhum estudo foi classificado globalmente com baixo risco de viés.

Os 12 estudos, foram ainda avaliados quanto à sua aplicabilidade. Neste contexto, 7 estudos (58,3%) apresentaram baixa preocupação, 1 estudo (8,3%) preocupação incerta e 4 estudos (33,3%) alta preocupação, principalmente devido a diferenças nas populações estudadas e contextos clínicos específicos.

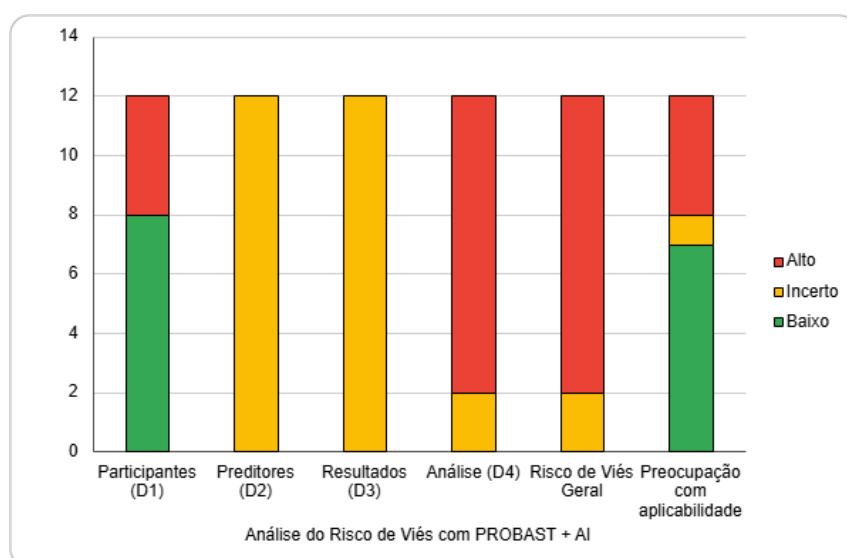


Figura 5.2: Gráfico da Análise do Risco de Vies com PROBAST+AI

Tabela 5.2: Resultado da aplicação do PROBAST+AI (Domínios- D1: Participantes, D2: Preditores, D3: resultados, D4: Análise)

Estudo	(D1)	(D2)	(D3)	(D4)	Risco de Viés Geral	Preocupação com Aplicabilidade
Liu et al. (2023b)	Baixo	Incerto	Incerto	Alto	Alto	Baixa
Chen et al. (2023)	Alto	Incerto	Incerto	Alto	Alto	Incerta
Liu et al. (2022)	Alto	Incerto	Incerto	Alto	Alto	Baixa
Zhang et al. (2024)	Alto	Incerto	Incerto	Alto	Alto	Alta
Klemencsics et al. (2016)	Baixo	Incerto	Incerto	Alto	Alto	Baixa
Wang et al. (2021)	Alto	Incerto	Incerto	Alto	Alto	Alta
Xiong et al. (2022)	Baixo	Incerto	Incerto	Alto	Alto	Baixa
Qiu et al. (2024)	Baixo	Incerto	Incerto	Incerto	Incerto	Baixa
Lou et al. (2025)	Baixo	Incerto	Incerto	Alto	Alto	Baixa
Lian et al. (2023)	Baixo	Incerto	Incerto	Incerto	Incerto	Baixa
Liu et al. (2023a)	Baixo	Incerto	Incerto	Alto	Alto	Alta
An et al. (2024)	Baixo	Incerto	Incerto	Alto	Alto	Alta

5.3 Características dos estudos

Os 12 estudos incluídos nesta revisão sistemática, foram publicados entre 2016 e Janeiro de 2025, sendo que a maioria, 11 estudos (91,7%), foram publicados entre os anos 2021-2025. Geograficamente, denota-se uma predominância de estudos asiáticos, 11 dos 12 estudos incluídos foram conduzidos na China, com apenas um estudo, o de Klemencsics et al. (2016), realizado na Hungria. No entanto, todos os estudos foram conduzidos, cada um num único hospital.

Relativamente ao desenho metodológico, 11 estudos adotaram uma abordagem retrospectiva, sendo que apenas um, o de Klemencsics et al. (2016), utilizou um desenho prospetivo. Quanto ao tipo de modelo preditivo desenvolvido, seguindo a abordagem metodológica adotada, oito estudos (66,7%) desenvolveram modelos baseados em regressão logística - sendo sete nomogramas com interface gráfica e um modelo apresentando apenas a equação matemática - enquanto quatro estudos (33,3%) realizaram estudos comparativos de *machine learning*, testando múltiplos algoritmos para seleção do modelo com melhor performance.

É importante notar que alguns estudos de nomogramas incorporaram técnicas de *machine learning* especificamente para seleção de variáveis preditivas, mantendo, contudo, a regressão logística como algoritmo final de previsão. Esta distinção metodológica reflete diferentes abordagens: modelos com algoritmo pré-definido versus seleção algorítmica baseada em comparação de performance.

A população total estudada foi de 9235 pacientes adultos, que teriam sido submetidos a diferentes tipos de cirurgia da coluna lombar. O tamanho da amostra variou consideravelmente entre estudos, desde 288 pacientes no estudo de Liu et al. (2022) até 1327 pacientes no estudo de Chen et al. (2023). Sendo que, a mediana de pacientes por estudo é de 714. A distribuição por género mostrou-se relativamente equilibrada na maioria dos estudos, com percentagens de participantes do sexo masculino a variar entre 41,3% e 56,9%.

Já as taxas de infeção do local cirúrgico apresentaram uma variação entre 2.9% a 50%. Esta

heterogeneidade nas taxas de infecção reflete principalmente diferenças nas populações estudadas, tipos de procedimentos cirúrgicos e complexidade dos casos. O estudo com a taxa mais elevada (50%) utilizou um desenho com grupos balanceados, selecionando intencionalmente igual número de casos com e sem ILC para otimizar o desenvolvimento do modelo de ML, não representando a incidência real na população. De notar, ainda que todos os estudos utilizaram os critérios de diagnóstico de ILC estabelecidos pelo CDC, garantindo consistência na definição do *outcome*.

Quanto aos tipos de procedimentos cirúrgicos, a maioria dos estudos (58,3%) incluiu procedimentos de fusão lombar, particularmente TLIF (Transforaminal Lumbar Interbody Fusion) e PLIF (Posterior Lumbar Interbody Fusion). Outros estudos abordaram cirurgias degenerativas lombares em geral, ou procedimentos minimamente invasivos, como o MIS-TLIF (Minimally Invasive Lumbar Interbody Fusion). Ainda, dois dos estudos focaram-se em populações específicas, um em pacientes diabéticos tipo 2 e outro em pacientes idosos (≥ 65 anos).

Os períodos de recolha de dados variaram entre 3 e 9 anos. Sendo que a maioria dos estudos a cobre períodos entre 2017 e 2022. Esta distribuição temporal permite uma análise representativa das práticas cirúrgicas contemporâneas em cirurgia de coluna lombar.

Para facilitar a análise comparativa dos modelos preditivos identificados, os dados extraídos foram organizados em três tabelas segundo o tipo de abordagem metodológica. A Tabela 5.3 apresenta estudos que desenvolveram nomogramas baseados em regressão logística, a Tabela 5.4 as características do estudo baseado em regressão logística, com equação matemática. E por último a tabela 5.5, que evidencia as características dos estudos de *Machine Learning*.

Tabela 5.3: Características dos estudos que desenvolveram nomogramas baseados em regressão logística

E	Autor (Ano)	País (Período)	Desenho	População	N Total		ILC (%)	Divisão dados (T/V) (%)	Modelo estatístico	Variável alvo	Preditores (n)	EPV	Métricas	Validação
					M (%)	F (%)								
1	Liu et al. (2023b)	China (01/18 - 06/21)	Retrospectivo	Pacientes ≥18 anos com doença degenerativa lombar submetidos a TLIF	1277		30 (3,1%) (coorte treino)	75/25	Regressão logística	ILC vs não-ILC	Tabagismo, diabetes, perda sanguínea intraoperatória, classe ASA, cálcio sérico, albumina, glicose sérica (7)	4,3	AUC: 0.988 Calibração: Excelente DCA: Net benefit positivo	Interna: divisão 75/25
2	Chen et al. (2023)	China (06/12 - 02/21)	Retrospectivo	Pacientes ≥18 anos submetidos a cirurgia de fixação interna lombar	1327		54 (4,1%)	70/30	RL + variáveis selecionadas por ML (LASSO+SVM-RFE+Random Forest)	ILC vs não-ILC	Alterações Modic, Espessura da camada de tecido adiposo subcutâneo, hemoglobina (Hb) e glicose (4)	9,5	AUC: 0.987 C index: 0.982	Interna: divisão 70/30
8	Qiu et al. (2024)	China (01/20 - 12/23)	Retrospectivo	Pacientes com hérnia discal lombar submetidos a cirurgia	723		241 (33,3%)	70/30	Regressão logística + LASSO	ILC vs não-ILC	Idade, IMC, Volume de drenagem pós-operatória, Esponja gelatina, Ausência antibiótico pré-op, Tempo trombina (6)	27,5	AUC: 0,782 Sensibilidade: 0,605 Especificidade: 0,823 Calibração: 0,174 Hosmer-Lemeshow: p=0,583	Interna: B=1000 e divisão 70/30
9	Lou et al. (2025)	China (01/19 - 12/22)	Retrospectivo	Pacientes adultos (≥18 anos) submetidos a cirurgia de fusão lombar posterior	1015		51 (5,02%)	-	Regressão logística	ILC vs não-ILC	Tabagismo, diabetes, duração da cirurgia, perda sanguínea, ASA ≥ 3, duração da drenagem, comprimento da incisão, IMC, fixação interna (9)	5,67	C-index: 0.779; AUC: 0.845; Calibração: Boa calibração visual	Interna: Bootstrap
10	Lian et al. (2023)	China (01/19 - 06/22)	Retrospectivo	Pacientes submetidos a TLIF por doença degenerativa lombar	383		40 (10,4%)	75/25	Regressão logística	ILC vs não-ILC	infiltração gordurosa LPM, diabetes, duração cirurgia (3)	10	C-index: 0,955; AUC: 0,929; Calibração: boa concordância DCA: 0.74	Interna: divisão 75/25
11	Liu et	China	Retrospe	Pacientes com	339		56	70/30	Regressão	ILC vs	FPG-CV (coeficiente de	4,9	AUC: 0,890;	Interna:

	al. (2023a)	(01/18 -- 09/21)	ctivo	diabetes tipo 2 submetidos a TLIF	43,7%	56,3%	(16,5%)		logística	não-ILC	variação da glicemia em jejum) pré-op - MFBG (glicemia média em jejum) pré-op - Duração do dreno - FPG-CV pós op - Esquema de controle glicêmico pré-operatório - Duração da diabetes (>5 anos) - N° de vértebras fundidas (≥2) (7)		Calibração: Excelente; DCA: Benefício clínico	divisão 70/30
12	An et al. (2024)	China (01/17 - 12/22)	Retrospe ctivo	Pacientes idosos (≥65 anos) submetidos a cirurgia aberta da coluna lombar	578		17 (2.9%)	-	LASSO + Regressão Logística	ILC vs não-ILC	Obesidade, hipoalbuminemia, história de álcool (3)	5,7	AUC: 0.879; Calibração: excelente; DCA: Benefício clínico quando probabilidade limiar >1% e <89%	Interna:Bootst rap 1000
				42%	58%									

Tabela 5.4 Características do estudo baseado em regressão logística, com equação matemática

E	Autor (Ano)	País (Período)	Desenho	População	N Total		ILC (%)	Divisão dados (T/V) (%)	Modelo estatístico	Variável alvo	Preditores (n)	EPV	Métricas	Validação
					M (%)	F (%)								
5	Klemencsics et al. (2016)	Hungria (05/09 - 12/12)	Prospectivo	Pacientes ≥18 anos submetidos a cirurgia lombar degenerativa eletiva	1030		37 (3,6%)	70/30	Regressão logística	ILC vs não-ILC	Idade, IMC, diabetes, doença cardíaca isquêmica, arritmia, doença hepática crônica, doença autoimune (7)	3,75	AUC: 0.72	Temporal
					43,1%	56,9%								

Tabela 5.5 Características dos estudos de Machine Learning

E	Autor (Ano)	País (Período)	Desenho	N Total		ILC (%)	Tipo cx	Algoritmos Testados	Melhor Algoritmo	Variável alvo	Preditores (n)	EPV	Divisão dados (T/V %)	Métricas	Validação	Software
				M (%)	F (%)											
3	Liu et al. (2022)	China (12/10 - 12/19)	Retrospectivo	288		144 (50%)	Coluna lombar posterior	LR, MLP, DT, RF, GBM, XGB	Extreme Gradient Boosting (XGB)	ILC vs não-ILC	Idade, diabetes mellitus, hipertensão, IMC, história cirúrgica prévia, duração cirurgia, número de segmentos fundidos, perda sanguínea, procedimento cirúrgico (9)	11,2	70/30	Sensibilidade: 0.834 Acurácia: 0.860 Precisão: 0.900 AUC: 0.923 F1-score: 0.864	Interna: (10-fold CV + test set)	Python
				-	-											
4	Zhang et al. (2024)	China (01/15 - 10/22)	Prospectivo	986		92 (9,33%)	MIS-TLIF, TLIF, PLIF	SVM, LR, RF, XGB, DT, KNN, NB	Naïve Bayes (NB)	ILC vs não-ILC	Idade, IMC, tabagismo, fístula de LCR, duração da drenagem e o nível de albumina pré-operatório (6)	12,3	80/20	Sensibilidade: 0.78 Especificidade: 0.88 Acurácia: 0.87 AUC: 0.96	Interna (5-fold cross-validation)	Python
				407 (41.3%)	579 (58,7%)											
6	Wang et al. (2021)	China (05/12 - 10/19)	Retrospectivo	705		33 (4,7%)	MIS-TLIF	KNN, DT, SVM, RF, MLP, NB	Naive Bayes (NB)	ILC vs não-ILC	HbA1c pré-operatória, perda sanguínea estimada, tempo operatório, albumina pré-operatória, IMC, idade (6)	5	90/10	AUC 0.78, Sensibilidade 0.93, Especificidade 0.82, Acurácia 0.90	Interna (10-fold cross-validation)	Python
				401 (56,9%)	304 (43.1%)											
7	Xiong et al. (2022)	China (01/19 - 08/21)	Retrospectivo	584		33 (5,7%)	PLIF	AdaBoost, LogitBoost, XGB, GBM, GLM, BCT, RF	AdaBoost Classification Trees	ILC vs não-ILC	Albumina pré-operatória, Diabetes, Rutura dural intraoperatória, História de doença reumática (4)	4,25	50/50	Sensibilidade: 0.9375 Precisão: 0.2308 Acurácia: 0.8247 F3 Score: 0.7177 AUC: 0.906 (com SMOTE)	Interna (Hold-out)	R Software
				263 (45.03 %)	321 (54.97 %)											

Legenda: Logistic Regression (LR), Multilayer Perceptron (MLP), Decision Tree (DT), Random Forest (RF), Gradient Boosting Machine (GBM), Extreme Gradient Boosting (XGB), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Naïve Bayes (NB), AdaBoost Classification Trees (AdaBoost), Boosted Logistic Regression (LogitBoost), Generalized Linear Model (GLM) e Boosted Classification Trees (BCT)

5.4 Modelos Preditivos

Embora exista debate na literatura sobre a classificação da regressão logística como método de *machine learning*, diversos estudos têm distinguido entre abordagens estatísticas tradicionais e técnicas avançadas de *machine learning* em modelos de previsão médica. Por exemplo, Song et al. (2023) compararam “regressão logística tradicional” com “modelos de *machine learning*” incluindo métodos ensemble como *Random Forest*, *XGBoost* e *AdaBoost*. Esta distinção demonstra a prática estabelecida de categorizar estas abordagens separadamente devido às suas diferentes complexidades algorítmicas e capacidades preditivas.

Para este estudo, e de forma a facilitar a comparação entre abordagens metodológicas, os resultados foram organizados em dois subcapítulos distintos.

O primeiro subcapítulo aborda modelos baseados em regressão logística, que podem incorporar outras técnicas de *machine learning* para seleção de variáveis preditivas, mas utilizam fundamentalmente a regressão logística como algoritmo de previsão final. Destes, sete estudos desenvolveram nomogramas como interface gráfica para aplicação clínica, enquanto um estudo apresentou apenas a equação matemática.

O segundo subcapítulo analisa modelos que testaram múltiplos algoritmos de *machine learning* - incluindo *Logistic Regression*, *Multilayer Perceptron*, *Decision Tree*, *Random Forest*, *Gradient Boosting Machine*, e *Extreme Gradient Boosting* - selecionando o algoritmo com melhor performance para o modelo final. Esta abordagem distingue-se pela comparação sistemática entre diferentes algoritmos e seleção baseada em métricas de desempenho

5.4.1 Modelos Baseados em Regressão Logística

Dos estudos analisados, oito (66,7%) desenvolveram modelos baseados em modelos de regressão logística para previsão de ILC em cirurgia de coluna lombar. Destes, sete desenvolveram nomogramas com interface gráfica e um apresentou apenas a equação matemática.

5.4.1.1 Nomogramas

Sete estudos desenvolveram nomogramas baseados em regressão logística. Todos seguiram um desenho retrospectivo e foram conduzidos na China.

A população total abrangida pelos estudos de nomogramas foi de 5642 pacientes. O tamanho amostral apresentou uma grande variação, entre 339 pacientes no estudo de Liu et al. (2023a) até 1.327 no estudo de Chen et al. (2023). Também a taxa de ILC oscilou entre 2,9%, no estudo de An et al. (2024), focado em pacientes idosos, a 33,3%, no estudo de Qiu et al. (2024).

A preferência pelos procedimentos de fusão lombar foi evidente, com particular destaque para a técnica TLIF.

Dois dos sete estudos desenvolveram modelos para populações clínicas específicas de alto risco. O que demonstra uma abordagem personalizada na previsão de ILC. Liu et al. (2023a) desenvolveram um nomograma exclusivamente para pacientes com diabetes tipo 2 submetidos a TLIF, incorporando variáveis

específicas como coeficiente de variação da glicemia em jejum (FPG-CV) e esquemas de controlo glicémico pré-operatório. Por sua vez, An et al. (2024), optaram por uma abordagem diferente, orientando o modelo para população geriátrica (≥ 65 anos) submetida a cirurgia aberta da coluna lombar.

Relativamente à metodologia empregue, identificaram-se duas abordagens principais. Cinco estudos utilizaram regressão logística convencional, enquanto três estudos adotaram estratégias híbridas, incorporando técnicas de *machine learning* para seleção de variáveis seguidas de regressão logística para a criação do modelo final.

O estudo de Chen et al. (2023) demonstrou a abordagem metodológica mais elaborada. O mesmo combinou três técnicas de *machine learning* para seleção de variáveis: LASSO (Least Absolute Shrinkage and Selection Operator), SVM-RFE (Support Vector Machine - Recursive Feature Elimination) e Random Forest. Após intersecção dos resultados, aplicaram regressão logística para construção do nomograma final.

O estudo de Qiu et al. (2024) combinou também regressão logística com LASSO, seguida de validação cruzada com *ten-fold cross-validation* para seleção das variáveis com maior correlação. An et al. (2024) também utilizou LASSO seguido de regressão logística.

Relativamente as variáveis preditivas incluídas nos modelos, as mesmas variam entre 3 a 9 preditores por estudo. As variáveis incorporadas com mais frequência foram: diabetes mellitus, IMC / obesidade, tabagismo, perda sanguínea intraoperatória / duração da cirurgia, Classe ASA, Albumina / hipoalbuminemia, idade e Drenagem pós-operatória / duração do dreno.

O rácio eventos por variável (EPV), demonstrou uma oscilação entre 4,3 no estudo de Liu et al. (2023b) e 27,5 de Qiu et al. (2024). Sendo que a maioria dos estudos apresentou EPV inferior a 10.

No que diz respeito à divisão dos dados, a mesma seguiu predominantemente uma proporção de 70/30 ou 75/25 para treino e validação, tendo apenas um estudo não especificando claramente esta divisão. Quanto aos métodos de validação, os sete estudos realizaram validação interna através de estratégias diversificadas como divisão dos dados em conjuntos de treino e validação, bootstrap com 1000 reamostragens, e validação cruzada. Nenhum dos estudos realizou validação externa.

Em relação às métricas de desempenho, os nomogramas demonstraram uma performance variável, mas globalmente satisfatória. A área sob a curva (AUC) variou entre 0,78 e 0,988, com a maioria dos estudos a reportar valores superiores a 0,80. Os modelos com melhor performance foram os de Liu et al. (2023b) e Chen et al. (2023), com AUC de 0,988 e 0,987, respetivamente. Em contraste, o estudo de Qiu et al. (2024) evidenciou desempenho preditivo razoável com AUC de 0,782, tendo complementado a análise com métricas adicionais como sensibilidade (0,605) e especificidade (0,823).

Cinco dos estudos procederam à avaliação da calibração dos seus modelos. Estes reportaram uma concordância excelente ou boa entre as probabilidades preditas pelos nomogramas e aquelas efetivamente observadas na prática clínica. Além disso, a análise de decisão clínica (DCA - Decision Curve Analysis) foi também reportada em quatro estudos, demonstrando benefício clínico líquido positivo dos nomogramas em diferentes limiares de probabilidade. Sendo que o estudo de Qiu et al. (2024) demonstrou que o nomograma apresenta benefício clínico quando o limiar do modelo é estabelecido entre 1% e 90%.

5.4.1.2 Modelo com Equação Matemática

O estudo de Klemencsics et al. (2016) foi o único a apresentar apenas a equação matemática do modelo de

regressão logística, sem interface gráfica de nomograma. Este estudo, realizado na Hungria (único não asiático), distingue-se também por ser o único de natureza prospetiva, abrangendo 1.030 pacientes.

Apresentou um AUC de 0,72 e um EPV de 3,75. O mesmo realizou validação temporal através de duas coortes consecutivas da mesma instituição, com 723 doentes para desenvolvimento do modelo (2009-2012) e 307 doentes para validação (2012). Contudo, não realizou validação externa.

5.4.2 Modelos de Machine Learning

Dos doze estudos analisados, quatro deles (33,3%) desenvolveram modelos preditivos utilizando algoritmos de *machine learning* para previsão de ILC em cirurgia de coluna lombar.

Os quatro estudos abrangeram uma população total de 2563 pacientes. Os tamanhos amostrais variaram entre 288, no estudo de Liu et al. (2022) e 986 pacientes no estudo de Zhang et al. (2024).

Quanto às taxas de ILC, as mesmas apresentaram maior variabilidade comparativamente aos estudos de nomogramas, oscilando entre 4,7% no estudo de Wang et al. (2021) e 50% no estudo de Liu et al. (2022). É importante notar que este último utilizou um desenho balanceado, selecionando intencionalmente igual número de casos com e sem ILC. Esta escolha foi realizada por forma a otimizar o desenvolvimento do modelo, não representando a incidência real na população.

Quanto aos tipos de procedimentos cirúrgicos, observou-se uma maior diversidade. Dois estudos focaram-se em técnicas minimamente invasivas (MIS-TLIF), enquanto os restantes abordaram procedimentos de fusão lombar posterior (PLIF) e cirurgias lombares posteriores em geral.

Os estudos demonstraram uma abordagem metodológica robusta, testando múltiplos algoritmos de *machine learning* antes de selecionar o modelo com melhor performance. Liu et al. (2022) testaram seis algoritmos diferentes, incluindo *Logistic Regression*, *Multilayer Perceptron*, *Decision Tree*, *Random Forest*, *Gradient Boosting Machine* (GBM) e *Extreme Gradient Boosting* (XGB), sendo que este último apresentou o melhor desempenho com AUC de 0,923. Por outro lado, também Wang et al. (2021) testaram seis algoritmos. No entanto denotam-se algumas diferenças, em vez de *Logistic Regression*, GBM e XGB, foram incluídos *K-Nearest Neighbors* (KNN), *Support Vector Machine* (SVM) e *Naïve Bayes*. Entre esses, *Naïve Bayes* foi selecionado como modelo final, com AUC de 0,78.

Zhang et al. (2024) avaliaram sete algoritmos, incluindo *Support Vector Machines*, *Naïve Bayes*, *Logistic Regression*, *Random Forest*, *XGBoost*, *Decision Tree* e *k-Nearest Neighbor*. Também neste estudo, o *Naïve Bayes* apresentou a melhor performance, com AUC de 0,96, sensibilidade de 0,78 e especificidade de 0,88. Já Xiong et al. (2022) testaram também sete algoritmos, porém diferentes: *Boosted Classification Trees*, *Boosted Logistic Regression*, *Extreme Gradient Boosting*, *Stochastic Gradient Boosting*, *Generalized Linear Model*, *Random Forest* e *AdaBoost Classification Trees* selecionando este último como modelo final.

A análise comparativa dos quatro estudos revelou padrões consistentes na performance dos diferentes tipos de algoritmos. Os métodos ensemble, particularmente *Extreme Gradient Boosting* e *AdaBoost Classification Trees*, demonstraram performance superior em metade dos estudos. No estudo de Liu et al. (2022), o XGB superou significativamente outros algoritmos com AUC de 0,923, comparado com 0,780 do *Decision Tree* e 0,759 do *Multilayer Perceptron*.

Também, os algoritmos Naïve Bayes mostraram-se surpreendentemente eficazes em dois estudos, alcançando as melhores performances em Zhang et al. (2024) e Wang et al. (2021), com AUCs de 0,96 e 0,78, respetivamente.

Destaca-se ainda o estudo de Xiong et al. (2022) pela utilização inovadora da técnica Synthetic Minority Oversampling Technique (SMOTE). Técnica usada para abordar o desequilíbrio de classes. Os mesmos reportaram que o AdaBoost Classification Trees sem SMOTE apresentou AUC de 0,8726, mas com sensibilidade de 0,6250, tendo significativamente melhorado com a aplicação de SMOTE, alcançando AUC de 0,906 e sensibilidade de 0,9375.

As variáveis preditivas incorporadas nos modelos de *machine learning* variaram entre 4 e 9 preditores por estudo. As variáveis mais frequentemente utilizadas foram idade, IMC, diabetes mellitus e duração da cirurgia. Alguns estudos incorporaram variáveis mais específicas, como marcadores laboratoriais (albumina, hemoglobina) e características técnicas da cirurgia. O rácio eventos por variável variou entre 4,25 e 12,3 nos estudos de Xiong et al. (2022) e Zhang et al. (2024), respetivamente. Sendo que dois estudos (50%) apresentaram um EPV inferior a 10, como observado nos nomogramas.

Em termos de desempenho, os modelos de *machine learning* demonstraram métricas variáveis, mas geralmente satisfatórias. A AUC variou entre 0,78 e 0,96, com a maioria dos estudos reportando valores superiores a 0,90. A sensibilidade variou entre 0,78 e 0,9375, enquanto a especificidade oscilou entre 0,82 e 0,88 nos dois estudos que a reportaram. Dois estudos avaliaram a calibração dos modelos, reportando boa calibração. Métricas adicionais como precisão, F1-score e acurácia foram reportadas em alguns estudos, demonstrando performance consistente.

Todos os estudos realizaram validação interna utilizando técnicas robustas. Dois estudos utilizaram 10-fold cross-validation, Zhang et al. (2024) implementou o 5-fold cross-validation, enquanto Xiong et al. (2022) utilizou hold-out validation com divisão 50/50. Nenhum dos estudos realizou validação externa, consistente com o observado nos estudos de nomogramas.

5.5 Variáveis preditivas

A análise dos 12 estudos incluídos identificou um total de 39 variáveis preditivas distintas incorporadas nos modelos, após agrupamento de variáveis similares. O número de variáveis por modelo variou entre 3 e 9 preditores, com uma mediana de 6 variáveis por estudo.

As variáveis preditivas foram categorizadas em seis grupos principais: demográficas, comorbilidades, laboratoriais, cirúrgicas, pós-operatórias e específicas.

Diabetes mellitus e IMC/obesidade foram as variáveis mais frequentemente incluídas, presentes em 7 estudos cada (58,3%). Idade foi incorporada em 5 estudos (41,7%).

Entre as comorbilidades, o tabagismo foi incluído em 3 estudos (25,0%) e a classificação ASA em 2 estudos (16,7%).

Relativamente aos marcadores laboratoriais, albumina sérica/hipoalbuminemia foi a variável mais frequente desta categoria, presente em 5 estudos (41,7%). Enquanto a glicose sérica foi incluída em 2 estudos

(16,7%) e hemoglobina em 1 estudo (8,3%).

As variáveis cirúrgicas mais comuns foram duração da cirurgia e perda sanguínea intraoperatória, incluídas em 4 estudos cada (33,3%).

Entre os fatores pós-operatórios, variáveis relacionadas com drenagem foram incluídas em 4 estudos (33,3%), sendo duração da drenagem a mais frequente em 3 estudos (25,0%).

Alguns estudos incorporaram variáveis específicas para populações particulares. Liu et al. (2023a) incluíram variáveis relacionadas com o controlo glicémico em pacientes diabéticos, enquanto Chen et al. (2023) incorporaram alterações de Modic e espessura do tecido adiposo subcutâneo como preditores específicos.

As variáveis cirúrgicas representaram a categoria mais numerosa (25,6%), seguidas das comorbilidades e laboratoriais (23,1% cada), demonstrando a importância tanto de fatores intrínsecos do paciente como de aspetos técnicos do procedimento cirúrgico.



Discussão

6. Discussão

Esta revisão sistemática fornece uma análise abrangente do impacto dos modelos preditivos para infecções do local cirúrgico em cirurgias de coluna lombar. Esta disponibiliza ainda, evidência importante sobre a eficácia, o potencial e principais limitações destes modelos na prática clínica contemporânea.

6.1 Qualidade Metodológica e Risco de Viés

A avaliação da qualidade metodológica dos estudos incluídos revelou uma conclusão preocupante: nenhum dos estudos analisados apresentou um baixo risco de viés global. Esta descoberta contrasta com outra revisão sistemática nesta mesma área, a de Lauinger et al. (2024), intitulada de “*Prediction models for risk assessment of surgical site infection after spinal surgery: A systematic review*”. Esta revisão utilizou a ferramenta QUIPS, para análise do risco de viés, e os autores identificaram que 74% dos estudos apresentaram baixo risco de viés (Lauinger et al., 2024), enquanto na presente revisão, aplicando o PROBAST+AI, 83,3% dos estudos foram classificados como alto risco.

Esta discrepância pode ser explicada pelas diferenças fundamentais entre as ferramentas utilizadas. A ferramenta QUIPS foi criada especificamente para avaliar o risco de viés em estudos de fatores prognósticos, onde o objetivo da modelação multivariável é identificar preditores associados ao *outcome*, sem o propósito de desenvolver uma ferramenta para previsões individualizadas (Moons et al., 2019). Por outro lado, o PROBAST analisa a qualidade metodológica de estudos que desenvolvem, validam ou atualizam modelos de previsão multivariáveis destinados a fazer previsões individualizadas (Moons et al., 2019).

Agrava estas limitações metodológicas o facto de 91,7% dos estudos (11/12) adotarem um desenho retrospectivo. Como explicado por Moons et al. (2019) no artigo de desenvolvimento do PROBAST, em estudos com desenho de coorte prospetivo, o cegamento dos avaliadores de preditores relativamente à informação sobre *outcomes* surge naturalmente. Contudo, em estudos retrospectivos, a ocorrência desse viés é mais provável, uma vez que os preditores e *outcomes* são avaliados dentro de um período temporal semelhante (Moons et al., 2019).

Esta explicação elucida precisamente porque 100% dos estudos foram classificados como risco incerto no domínio D3 (*outcomes*) e domínio D2 (preditores) do PROBAST+AI. O desenho retrospectivo inerentemente compromete o cegamento, criando potencial para viés de informação.

6.2 Concentração geográfica e limitações de generalização

Além das limitações metodológicas identificadas, a distribuição geográfica dos estudos levanta questões adicionais sobre a generalização dos resultados.

Uma conclusão notável desta revisão foi a extrema concentração geográfica dos estudos, com 91,7% (11/12) conduzidos na China. Também este padrão contrasta marcadamente com a literatura internacional, onde Lauinger et al. (2024) reportaram uma distribuição mais equilibrada: 40% nos EUA, 23% na China e 8,5 % no Japão, os restantes foram distribuídos pela Europa e outros países. Essa predominância regional pode refletir o rápido desenvolvimento da investigação em inteligência artificial na China e possível viés de publicação em bases de dados internacionais.

Além disso, predomínio de estudos retrospectivos, aliados a esta concentração geográfica desigual, levanta questões importantes sobre a generalização dos resultados para diferentes contextos clínicos, sistemas de saúde e características populacionais. A validação externa geográfica é fundamental quando se questiona se um modelo preditivo desenvolvido numa região específica é transportável para outras populações (Ramspek et al., 2020). Esta limitação é reforçada pelo facto de que nenhum dos 11 estudos chineses incluídos realizou validação externa, restringindo ainda mais a aplicabilidade dos resultados a contextos internacionais.

Adicionalmente, as diferenças populacionais entre a China e outras regiões podem impactar significativamente a performance dos modelos. Por exemplo, diferenças na prevalência de obesidade ($IMC \geq 30 \text{ kg/m}^2$), em 2022, com a China a apresentar uma menor prevalência, de 8,3%, comparativamente aos EUA 42,0%, a Portugal, com 21,8%, Hungria com 31,7% e França com 9,7% (World Health Organization, 2022), podem comprometer a aplicabilidade de modelos que incorporam obesidade como preditor principal.

Em linha com esta observação, a variabilidade do impacto dos preditores pode influenciar significativamente a performance dos modelos quando aplicado a diferentes contextos e populações (Ramspek et al., 2020). Isto porque, o mesmo preditor pode ter uma importância prognóstica diferente em populações distintas, como acontece com o nível socioeconómico que pode ser um fator de risco importante em países com sistemas de saúde privatizados, mas menos preditivo em países com cuidados de saúde universais (Ramspek et al., 2020).

Tendo em conta que as variações na prestação de cuidados de saúde, demografia dos pacientes e particularidades locais irão ditar a performance de um modelo preditivo específico, torna-se fundamental a realização de validações externas (Collins et al., 2024). Ainda, no seu estudo, Tkatschenko et al. (2023), demonstraram que a nível mundial, existe uma heterogeneidade considerável nas estratégias de prevenção de ILC em cirurgia de coluna. Os mesmos evidenciam grandes variações nas medidas pré-, peri- e pós-operatórias para prevenir ILC (Tkatschenko et al., 2023). Esta variação nas práticas clínicas pode impactar significativamente a performance dos modelos preditivos quando aplicados em diferentes contextos geográficos (Collins et al., 2024).

Portanto, esta concentração geográfica, aliada à ausência de validação externa nos estudos chineses, representa uma limitação significativa para a implementação clínica destes modelos preditivos. A predominância de estudos numa única região geográfica, com práticas clínicas e características populacionais específicas, compromete a confiança na generalização destes resultados para outros contextos de saúde.

Para mitigar estas limitações, recomenda-se uma abordagem estruturada de validação externa. Primeiro, validação temporal em centros chineses para confirmar estabilidade dos modelos. Segundo, validação geográfica

em centros de diferentes regiões (Europa, América do Norte, outras regiões asiáticas). Terceiro, se a performance se degradar significativamente, considerar re-calibração ou re-desenvolvimento dos modelos incorporando dados locais.

Assim, antes da implementação clínica internacional, será essencial proceder à validação externa destes modelos em populações diversificadas, considerando as especificidades locais dos sistemas de saúde, protocolos cirúrgicos e características demográficas. Desta forma, aumenta-se a probabilidade de que os modelos preditivos se tornem efetivamente úteis e aplicáveis, contribuindo para a melhoria dos cuidados em cirurgia de coluna lombar.

6.3 Performance Preditiva

Além das limitações geográficas, a análise da performance preditiva dos modelos revela inconsistências adicionais que comprometem sua aplicabilidade clínica.

A distinção metodológica adotada nesta revisão reflete diferenças fundamentais nas abordagens de desenvolvimento dos modelos. Os estudos classificados como baseados em regressão logística utilizaram este algoritmo como método final de previsão, independentemente de terem empregado técnicas de *machine learning* para seleção de variáveis. Em contraste, os estudos de *machine learning* caracterizaram-se pela comparação sistemática entre múltiplos algoritmos, selecionando o de melhor performance através de métricas objetivas. Esta distinção é metodologicamente relevante pois reflete diferentes filosofias de desenvolvimento: modelos com algoritmo pré-definido versus seleção algorítmica baseada em performance.

A maioria dos modelos demonstrou performance discriminativa (AUC) superior a 0,70. Segundo Mandrekar (2010) uma AUC de 0,5 sugere que o algoritmo não apresenta capacidade discriminatória, de 0,7 a 0,8 é considerada aceitável, 0,8 a 0,9 é considerada excelente e mais de 0,9 é considerada excepcional. Seguindo esta lógica, consideramos então que os modelos apresentaram performances aceitáveis a excepcionais.

Os valores de AUC, demonstram uma variabilidade grande, oscilando entre 0,72 e 0,988, - uma diferença de 0,268 pontos. Esta heterogeneidade é notável, considerando que todos os modelos visam prever a mesma condição clínica em populações aparentemente semelhantes. A distribuição dos valores AUC revela um padrão preocupante: enquanto a maioria dos estudos reportou valores superiores a 0,80, os valores excepcionalmente elevados (>0,95) levantam suspeitas metodológicas fundamentadas.

Os modelos com performance aparentemente superior foram os de Liu et al. (2023b) e Chen et al. (2023), com AUC de 0,988 e 0,987, respectivamente. Contudo, estes valores aproximam-se da discriminação perfeita e levantam questões sobre a sua plausibilidade clínica, especialmente considerando a complexidade multifatorial das infecções do local cirúrgico.

A análise revela uma correlação inversa preocupante entre adequação metodológica e performance reportada. Destes, o estudo de Liu et al. (2023b), apresenta o AUC mais elevado, de 0,988. No entanto, este valor elevado merece interpretação cuidadosa, dado que o estudo apresenta um EPV de apenas 4,3 (30 eventos / 7 variáveis preditoras). O que representa menos de metade do mínimo recomendado por Peduzzi et al. (1996), que demonstraram que, tradicionalmente, para valores de EPV de 10 ou superiores não ocorrem problemas

significativos na modelação de regressão logística.

Uma possível explicação metodológica para justificar estes valores de AUC elevados, que requer análise cuidadosa, pode ser o rácio eventos por variável baixo. Uma vez que, EPVs baixos aumentam substancialmente o risco de instabilidade do modelo e overfitting (Peduzzi et al., 1996).

Em contraste, o estudo de Qiu et al. (2024), com o EPV mais adequado (27,5), reportou um AUC mais conservador de 0,782, sugerindo uma estimativa mais realista da performance preditiva. Criticamente, 66,7% dos estudos apresentaram EPV inferior a 10, comprometendo a estabilidade e generalização dos modelos desenvolvidos.

Contudo, evidências contemporâneas questionam a adequação da regra $EPV \geq 10$ como critério único. Riley et al. (2019) propuseram critérios mais robustos que consideram simultaneamente o shrinkage factor ($\geq 0,9$), a diferença entre R^2 aparente e ajustado ($\leq 0,05$), e a estimação precisa do risco populacional. Em exemplos práticos, os autores demonstram que, por vezes, o EPV necessário foi superior a 20, ou seja, mais do dobro da regra tradicional (Riley et al., 2019). Isto sugere que a inadequação metodológica dos estudos analisados pode ser ainda mais severa.

Para além disso, a ausência de validação externa em 100% dos estudos impede completamente a avaliação da verdadeira capacidade de generalização dos modelos. Esta limitação é particularmente grave, pois valores de AUC obtidos apenas com validação interna tendem a superestimar sistematicamente a performance real dos modelos quando aplicados a populações independentes. O estudo de Klemencsics et al. (2016), embora tenha realizado validação temporal, apresentou um AUC de 0,72 - o valor mais baixo entre todos os estudos - sugerindo que validações internas mais rigorosas podem revelar limitações mascaradas por metodologias menos robustas.

Complementando estas limitações, a análise da validação interna revelou uma abordagem metodológica heterogénea entre os estudos incluídos. Embora tenham realizado alguma forma de validação interna, as estratégias variaram desde simples divisões treino/validação até métodos mais robustos como bootstrap e validação cruzada. Esta heterogeneidade, combinada com EPV inadequados e falta de validação externa, representa uma lacuna crítica que pode limitar severamente a aplicabilidade clínica dos modelos desenvolvidos.

Ainda, alguns estudos utilizaram desenhos metodológicos que podem inflacionar artificialmente a performance. Liu et al. (2022), com AUC de 0,923, aplicaram um desenho balanceado com 50% de casos de ILC, uma prevalência que não reflete a incidência real. Este balanceamento artificial pode inflacionar as métricas de performance e comprometer a generalização dos resultados para a prática clínica, onde a prevalência é substancialmente menor. Consequentemente, os valores preditivos positivos reportados podem ser irrealisticamente elevados, limitando a aplicabilidade clínica dos modelos desenvolvidos.

Contrariamente ao expectável, não se observou uma superioridade consistente dos algoritmos de *machine learning* sobre os nomogramas tradicionais. Esta descoberta alinha-se com evidência crescente de que, métodos estatísticos convencionais podem ser tão eficazes quanto algoritmos mais complexos (Christodoulou et al., 2019).

De notar que os algoritmos *ensemble* como Extreme Gradient Boosting e AdaBoost Classification Trees demonstraram performance superior em metade dos estudos de *machine learning*. Curiosamente, o algoritmo Naive Bayes destacou-se, tendo as melhores performances na outra metade dos estudos de ML: Zhang et al. (2024) com AUC de 0,96 e Wang et al. (2021) com AUC de 0,78.

Para além da discriminação (AUC), a calibração dos modelos é igualmente crucial para a aplicabilidade

clínica. Modelos bem calibrados garantem que as probabilidades preditas correspondem às frequências observadas, sendo essencial para a tomada de decisão clínica informada. Contudo, a calibração foi reportada em apenas seis estudos de nomogramas e dois de *machine learning*, representando uma lacuna importante na avaliação da qualidade dos modelos.

A análise de decisão clínica (DCA), reportada em quatro estudos, demonstrou benefício clínico líquido positivo dos nomogramas em diferentes limiares de probabilidade, sugerindo utilidade prática na tomada de decisão clínica.

6.4 Variáveis Preditivas

A análise das 39 variáveis preditivas identificadas revela consistência com os fatores de risco estabelecidos na literatura. As variáveis mais frequentemente incluídas foram: diabetes *mellitus* (58,3%), IMC/obesidade (58,3%), idade (41,7%) e albumina sérica (41,7%). Estas correspondem diretamente aos fatores pré-operatórios mais consistentemente identificados em investigações prévias: idade avançada, obesidade, diabetes *mellitus* e estado nutricional deficiente (Janssen et al., 2019; Lubelski et al., 2021).

Igualmente relevante é a inclusão frequente de variáveis intraoperatórias como duração da cirurgia e perda sanguínea (33,3% cada). Os mesmos foram também identificados na literatura por Janssen et al. (2019) e Lubelski et al. (2021). A variabilidade na sua inclusão pode refletir diferenças na disponibilidade de dados retrospectivos ou nas estratégias de seleção de variáveis adotadas pelos diferentes estudos. Esta variação pode também relacionar-se com o timing pretendido de aplicação do modelo: se pré-cirúrgico para apoio à decisão de intervenção, ou pós-cirúrgico para estratificação de risco no seguimento.

Esta convergência entre os preditores identificados nos modelos e os fatores de risco estabelecidos na literatura reforça a validade dos modelos desenvolvidos. Este facto sugere que, independentemente da abordagem metodológica utilizada, certas variáveis mantêm-se como preditores de ILC em cirurgia de coluna lombar. Destaca-se ainda a incorporação de variáveis específicas para populações especiais, como marcadores de controlo glicémico em diabéticos tipo 2 e fatores geriátricos em pacientes idosos, evidenciando uma abordagem direcionada para subgrupos de maior risco.

6.5 Implicações para a Prática Clínica

Apesar das limitações metodológicas identificadas, a maioria dos modelos demonstrou performance discriminativa aceitável ($AUC > 0,70$), sugerindo potencial clínico promissor. Contudo, várias considerações são essenciais antes da implementação clínica.

Dado o predomínio de estudos chineses e a ausência de validação externa na maioria dos casos, qualquer implementação deve ser obrigatoriamente precedida da validação externa rigorosa. Segundo Altman et al. (2009) não se deve implementar modelos que não tenham sido validados.

Esta validação do modelo, deverá não só confirmar a performance discriminativa, mas também garantir

calibração adequada no contexto clínico específico (Altman et al., 2009). Devem ser consideradas as diferenças populacionais, práticas cirúrgicas e sistemas de saúde locais.

Embora a calibração tenha sido reportada em mais de metade dos estudos, a ausência desta avaliação nos restantes representa uma lacuna importante. A calibração é fundamental para a aplicabilidade clínica, pois garante que as probabilidades fornecidas pelo modelo refletem o risco real do paciente, permitindo decisões clínicas informadas (Van Calster et al., 2019). Modelos mal calibrados podem levar a intervenções desnecessárias e criar falsas expectativas, tanto no paciente como no clínico (Van Calster et al., 2019). Importa referir ainda, que todas estas consequências podem também levar a um aumento dos custos hospitalares.

Estas conclusões corroboram a observação de Altman et al. (2009), que poucos modelos prognósticos são rotineiramente utilizados na prática clínica, provavelmente porque a maioria não foi externamente validada.

Para além da validação e calibração, outro aspeto crucial para a utilização clínica destes modelos é a forma como são integrados nos sistemas de saúde. Posto isto, e para maximizar a utilidade clínica, os modelos devem ser integrados de forma cuidadosa nos sistemas de registos eletrónicos de saúde. A implementação de modelos preditivos em sistemas EHR (Registos Eletrónicos de Saúde) não é direta, uma vez que é área emergente onde as melhores práticas ainda não estão bem estabelecidas (Lee et al., 2020). A integração eficaz requer consideração de múltiplos factores, incluindo explicabilidade dos algoritmos, recursos computacionais, escalabilidade, padronização de dados e integração significativa nos fluxos de trabalho clínicos (Lee et al., 2020).

Os desafios de implementação identificados por Lee et al. (2020) incluem fadiga de alertas, falta de formação, aumento da carga de trabalho da equipa de cuidados e interrupções no fluxo de trabalho.

Adicionalmente, a performance dos modelos deve ser monitorizada continuamente para detetar possível degradação ao longo do tempo, particularmente face a mudanças nas práticas clínicas ou características da população.

Finalmente, mas não menos importante, os clínicos necessitam de formação adequada para interpretar e utilizar corretamente as previsões dos modelos, garantindo que a informação preditiva seja efetivamente traduzida em decisões clínicas informadas. Como estabelecido por Paranjape et al. (2019), à medida que a prática da medicina entra na era da inteligência artificial, o uso de dados para melhorar a tomada de decisão clínica crescerá. Os profissionais precisam de ser adequadamente formados nesta nova tecnologia, nas suas vantagens para melhorar o custo, qualidade e acesso aos cuidados de saúde, e nas suas limitações, como transparência e responsabilidade (Paranjape et al., 2019).

Em síntese, embora os modelos demonstrem potencial clínico, a implementação segura requer uma abordagem sistemática que priorize validação local, integração cuidadosa, monitorização contínua e formação adequada dos profissionais de saúde.

6.6 Limitações desta Revisão

Esta revisão apresenta limitações que devem ser consideradas na interpretação dos resultados. A heterogeneidade metodológica dos estudos impediu a realização de meta-análise, limitando a síntese

quantitativa.

Além disso, a busca foi restrita a artigos em inglês e português, podendo ter excluído estudos relevantes noutros idiomas. Não foi também possível aceder ao texto completo de alguns artigos potencialmente relevantes, o que pode ter resultado na exclusão de estudos pertinentes. Ainda relativamente a pesquisa, uma vez que a mesma foi limitada a bases de dados específicas, pode não ter capturado toda a literatura disponível sobre o tema.

Uma limitação particularmente significativa é o facto de todos os estudos incluídos terem sido conduzidos em centros únicos. Esta abordagem unicêntrica pode não capturar a variabilidade nas práticas clínicas, recursos hospitalares e características populacionais que existem entre diferentes instituições. Esta limitação reforça ainda mais as preocupações sobre a generalização dos resultados para contextos clínicos mais amplos.

Além disso, a rápida evolução dos algoritmos de *machine learning* pode tornar alguns achados rapidamente desatualizados, particularmente no que se refere às técnicas mais avançadas de ensemble e deep learning. Esta dinâmica tecnológica representa um desafio contínuo para revisões sistemáticas nesta área.

Por fim, embora a avaliação da qualidade metodológica tenha sido realizada por dois revisores independentes com discussão para consenso, minimizando embora não eliminando completamente a subjetividade inerente a este processo, reconhece-se que a interpretação de alguns critérios do PROBAST+AI pode conter elementos de julgamento subjetivo.

Apesar destas limitações, considera-se que esta revisão fornece uma avaliação abrangente e rigorosa do estado atual da evidência sobre modelos preditivos para ILC em cirurgia de coluna lombar, identificando lacunas críticas que devem orientar investigações futuras.



Conclusão

7. Conclusão

Esta revisão sistemática representa a primeira análise abrangente dos modelos preditivos para infecções do local cirúrgico em cirurgias de coluna lombar, fornecendo evidência importante sobre a sua eficácia e limitações na prática clínica contemporânea.

Os resultados demonstram que tanto nomogramas quanto algoritmos de *machine learning* apresentam performance discriminativa satisfatória (AUC 0,72-0,988), sem superioridade consistente dos algoritmos mais complexos. As variáveis preditivas mais relevantes, como a diabetes mellitus, IMC/obesidade, idade e albumina sérica, alinham-se com os fatores de risco estabelecidos na literatura, reforçando a validade clínica dos modelos.

Contudo, a avaliação da qualidade metodológica revelou limitações críticas: nenhum estudo apresentou baixo risco de viés, 83,3% foram classificados como alto risco, e observou-se concentração geográfica extrema (91,7% na China) com ausência de validação externa. O rácio eventos por variável inadequado (EPV < 10 em 66,7% dos estudos) e a natureza unicêntrica de todos os estudos comprometem significativamente a generalização dos resultados.

Apesar do potencial clínico promissor, a implementação responsável requer obrigatoriamente a validação externa rigorosa em populações diversificadas. As prioridades futuras incluem estudos prospectivos multicêntricos com amostras adequadas, expansão geográfica da investigação e desenvolvimento de estratégias de implementação clínica.

Em síntese, os modelos preditivos demonstram potencial significativo para melhorar a avaliação de risco individualizada em cirurgia de coluna lombar. Contudo, apenas através de uma abordagem metodologicamente robusta e geograficamente diversificada será possível desenvolver modelos confiáveis que contribuam efetivamente para a melhoria dos cuidados de saúde e redução da incidência de ILC.



Recomendações e estudos futuros

8. Recomendações e estudos futuros

As evidências na literatura ainda são insuficientes acerca da eficácia dos modelos preditivos para ILC em cirurgia de coluna lombar na prática clínica real. A concentração geográfica extrema, a ausência completa de validação externa na maioria dos casos e o predomínio de estudos retrospectivos unicêntricos comprometem significativamente a generalização dos resultados para contextos clínicos internacionais.

É prioritário que sejam realizados estudos prospectivos multicêntricos com amostras representativas e adequadas. Neste contexto, estudos futuros devem não só, ter em conta a regra $EPV \geq 10$, mas também aplicar os critérios contemporâneos de Riley et al. (2019) que consideram simultaneamente o shrinkage factor, a diferença entre R^2 aparente e ajustado, e a estimação precisa do risco populacional para determinação rigorosa do tamanho amostral.

Ainda, estes estudos devem adotar desenhos metodologicamente robustos que incluam validação externa rigorosa, e aplicação de métricas de avaliação padronizadas que considerem não apenas a discriminação (AUC) mas também a calibração dos modelos. Estes estudos devem expandir-se geograficamente para populações não-asiáticas, considerando as variações nos sistemas de saúde, práticas clínicas e características demográficas que podem influenciar a performance dos modelos.

Além disso, encoraja-se que sejam conduzidos estudos de implementação clínica que avaliem o impacto real dos modelos preditivos na redução da incidência de ILC, acompanhados por períodos mais longos de seguimento para demonstrar efeitos mais significativos nos outcomes dos pacientes. A investigação sobre estratégias de integração fluida nos sistemas de informação hospitalar, formação adequada dos clínicos e desenvolvimento de interfaces user-friendly é fundamental para maximizar a adoção e utilização efetiva destes modelos na tomada de decisão clínica.

Finalmente, recomenda-se o desenvolvimento de abordagens híbridas que combinem a interpretabilidade dos nomogramas com as capacidades avançadas de *machine learning*, baseadas nas melhores práticas identificadas e nos modelos de maior sucesso. Isto para uma melhor aceitação e implementação dos sistemas de apoio à decisão clínica no manejo preventivo das ILC em cirurgia de coluna lombar.



Referências

9. Referências

- Alfin, D. J., Shilong, D. J., Bot, G. M., & Dengunu Salun, W. (2025). Surgical site infection rate in spine surgery, incidence, and risk factors: a ten-year retrospective cohort review in a developing neurosurgical centre. *BMC surgery*, 25(1), 127. <https://doi.org/10.1186/s12893-025-02846-4>
- Altman, D. G., Vergouwe, Y., Royston, P., & Moons, K. G. (2009). Prognosis and prognostic research: validating a prognostic model. *BMJ (Clinical research ed.)*, 338, b605. <https://doi.org/10.1136/bmj.b605>
- An, Y., Cui, X., Wang, H., Sun, Y., Zhu, B., Feng, S., & Jiang, J. (2024). Nomogram for predicting surgical site infections in elderly patients after open lumbar spine surgery: A retrospective study. *International wound journal*, 21(3), e14734. <https://doi.org/10.1111/iwj.14734>
- Atkinson, R. A., Jones, A., Ousey, K., & Stephenson, J. (2017). Management and cost of surgical site infection in patients undergoing surgery for spinal metastasis. *The Journal of hospital infection*, 95(2), 148–153. <https://doi.org/10.1016/j.jhin.2016.11.016>
- Balachandran, V. P., Gonen, M., Smith, J. J., & DeMatteo, R. P. (2015). Nomograms in oncology: more than meets the eye. *The Lancet. Oncology*, 16(4), e173–e180. [https://doi.org/10.1016/S1470-2045\(14\)71116-7](https://doi.org/10.1016/S1470-2045(14)71116-7)
- Borchardt, R. A., & Tzizik, D. (2018). Update on surgical site infections: The new CDC guidelines. *JAAPA : official journal of the American Academy of Physician Assistants*, 31(4), 52–54. <https://doi.org/10.1097/01.JAA.0000531052.82007.42>
- Chen, T., Liu, C., Zhang, Z., Liang, T., Zhu, J., Zhou, C., Wu, S., Yao, Y., Huang, C., Zhang, B., Feng, S., Wang, Z., Huang, S., Sun, X., Chen, L., & Zhan, X. (2023). Using Machine Learning to Predict Surgical Site Infection After Lumbar Spine Surgery. *Infection and drug resistance*, 16, 5197–5207. <https://doi.org/10.2147/IDR.S417431>
- Christodoulou, E., Ma, J., Collins, G. S., Steyerberg, E. W., Verbakel, J. Y., & Van Calster, B. (2019). A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of clinical epidemiology*, 110, 12–22. <https://doi.org/10.1016/j.jclinepi.2019.02.004>
- Collins, G. S., Dhiman, P., Ma, J., Schluskel, M. M., Archer, L., Van Calster, B., Harrell, F. E., Jr, Martin, G. P., Moons, K. G. M., van Smeden, M., Sperrin, M., Bullock, G. S., & Riley, R. D. (2024). Evaluation of clinical prediction models (part 1): from development to external validation. *BMJ (Clinical research ed.)*, 384, e074819. <https://doi.org/10.1136/bmj-2023-074819>
- Direção-Geral da Saúde. (2022, novembro 17). "Feixe de Intervenções" para a Prevenção da Infecção do Local Cirúrgico: Norma 020/2015 (atualizada). https://normas.dgs.min-saude.pt/wp-content/uploads/2015/12/norma_020_2015_atualizada_17_11_2022_prev_inf_local_cirurgico.pdf
- Dorow, M., Löbner, M., Stein, J., Konnopka, A., Meisel, H. J., Günther, L., Meixensberger, J., Stengler, K., König, H. H., & Riedel-Heller, S. G. (2017). Risk Factors for Postoperative Pain Intensity in Patients Undergoing Lumbar Disc Surgery: A Systematic Review. *PloS one*, 12(1), e0170303.

- <https://doi.org/10.1371/journal.pone.0170303>
- Ge, Z., Liu, X., Jing, X., Wang, J., Guo, Y., Yang, H., & Cui, X. (2024). Logistic regression model predicts early surgical site infection after spinal fusion: a retrospective cohort study. *Journal of Hospital Infection*, 149, 65-76.
- Gu, K., Lu, S. Machine learning-based predictive tools and nomogram for in-hospital mortality in critically ill cancer patients: development and external validation using retrospective cohorts. *BMC Med Inform Decis Mak* **25**, 251 (2025). <https://doi.org/10.1186/s12911-025-03054-z>
- Hopkins, B. S., Mazmudar, A., Driscoll, C., Svet, M., Goergen, J., Kelsten, M., Shlobin, N. A., Kesavabhotla, K., Smith, Z. A., & Dahdaleh, N. S. (2020). Using artificial intelligence (AI) to predict postoperative surgical site infection: A retrospective cohort of 4046 posterior spinal fusions. *Clinical Neurology and Neurosurgery*, 192, 105718. <https://doi.org/10.1016/j.clineuro.2020.105718>
- Iasonos, A., Schrag, D., Raj, G. V., & Panageas, K. S. (2008). How to build and interpret a nomogram for cancer prognosis. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, 26(8), 1364–1370. <https://doi.org/10.1200/JCO.2007.12.9791>
- Iwagami, M., & Matsui, H. (2022). Introduction to Clinical Prediction Models. *Annals of clinical epidemiology*, 4(3), 72–80. <https://doi.org/10.37737/ace.22010>
- Janssen, D. M., van Kuijk, S. M., d'Aumerie, B., & Willems, P. (2019). A prediction model of surgical site infection after instrumented thoracolumbar spine surgery in adults. *European Spine Journal*, 28, 775782.
- Kim, J. S., Merrill, R. K., Arvind, V., Kaji, D., Pasik, S. D., Nwachukwu, C. C., Vargas, L., Osman, N. S., Oermann, E. K., Caridi, J. M., & Cho, S. K. (2018). Examining the Ability of Artificial Neural Networks Machine Learning Models to Accurately Predict Complications Following Posterior Lumbar Spine Fusion. *Spine*, 43(12), 853–860. <https://doi.org/10.1097/BRS.0000000000002442>
- Klemencsics, I., Lazary, A., Szoverfi, Z., Bozsodi, A., Eltes, P., & Varga, P. P. (2016). Risk factors for surgical site infection in elective routine degenerative lumbar surgeries. *The spine journal : official journal of the North American Spine Society*, 16(11), 1377–1383. <https://doi.org/10.1016/j.spinee.2016.08.018>
- Lauinger, A. R., Blake, S., Fullenkamp, A., Polites, G., Grauer, J. N., & Arnold, P. M. (2024). Prediction models for risk assessment of surgical site infection after spinal surgery: A systematic review. *North American Spine Society journal*, 19, 100518. <https://doi.org/10.1016/j.xnsj.2024.100518>
- Lee, T. C., Shah, N. U., Haack, A., & Baxter, S. L. (2020). Clinical Implementation of Predictive Models Embedded within Electronic Health Record Systems: A Systematic Review. *Informatics (MDPI)*, 7(3), 25. <https://doi.org/10.3390/informatics7030025>
- Lian, J., Wang, Y., Yan, X., Xu, G., Jia, M., Yang, J., Ying, J., & Teng, H. (2023). Development and validation of a nomogram to predict the risk of surgical site infection within 1 month after transforaminal lumbar interbody fusion. *Journal of orthopaedic surgery and research*, 18(1), 105. <https://doi.org/10.1186/s13018-023-03550-w>
- Lim, S., Edelstein, A. I., Patel, A. A., Kim, B. D., & Kim, J. Y. S. (2018). Risk Factors for Postoperative Infections After Single-Level Lumbar Fusion Surgery. *Spine*, 43(3), 215–222. <https://doi.org/10.1097/BRS.0000000000000608>
- Liu, H., Zhang, W., Hu, Q., Liu, L., Xie, Z., Xu, Y., Jing, G., & Wang, Y. (2023a). A nomogram for accurately predicting the surgical site infection following transforaminal lumbar interbody fusion in type 2 diabetes patients, based on glycemic variability. *International wound journal*, 20(4), 981–994. <https://doi.org/10.1111/iwj.13948>
- Liu, H., Zhang, W., Zhang, Y., Zhang, S., Jin, G., & Li, X. (2023b). Establishment and validation of a nomogram model for postoperative surgical site infection after transforaminal lumbar interbody fusion: A retrospective

- observational study. *Surgery*, 174(5), 1220–1226. <https://doi.org/10.1016/j.surg.2023.07.019>
- Liu, W. C., Ying, H., Liao, W. J., Li, M. P., Zhang, Y., Luo, K., Sun, B. L., Liu, Z. L., & Liu, J. M. (2022). Using Preoperative and Intraoperative Factors to Predict the Risk of Surgical Site Infections After Lumbar Spinal Surgery: A Machine Learning-Based Study. *World neurosurgery*, 162, e553–e560. <https://doi.org/10.1016/j.wneu.2022.03.060>
- Lubelski, D., Feghali, J., Ehresman, J., Pennington, Z., Schilling, A., Huq, S., ... & Sciubba, D. M. (2021). Web-based calculator predicts surgical-site infection after thoracolumbar spine surgery. *World neurosurgery*, 151, e571–e578.
- Luo, J.-Z., Lin, J.-Z., Chen, Q.-F., Yang, C.-J., & Zhou, C.-S. (2025). Construction and validation of a nomogram predictive model for assessing the risk of surgical site infections following posterior lumbar fusion surgery. *Scientific Reports*, 15(1), 1023. <https://doi.org/10.1038/s41598-024-84174-w>
- Lupo, B. D., Jameson, W. P., Quinones, C. J., Malek, A. E., Kumbhare, D., Guthikonda, B., & Hoang, S. (2025). Risk Factors and Outcomes of Surgical Site Infections of the Spine: A Retrospective Multi-Center Analysis. *Journal of clinical medicine*, 14(10), 3520. <https://doi.org/10.3390/jcm14103520>
- Mandrekar J. N. (2010). Receiver operating characteristic curve in diagnostic test assessment. *Journal of thoracic oncology : official publication of the International Association for the Study of Lung Cancer*, 5(9), 1315–1316. <https://doi.org/10.1097/JTO.0b013e3181ec173d>
- Matsumoto, H., Franzone, J. M., Sinha, R., Roye, B. D., Glotzbeker, M. P., Skaggs, D. L., Flynn, J. M., Lenke, L. G., Sponseller, P. D., & Vitale, M. G. (2022). A novel risk calculator predicting surgical site infection after spinal surgery in patients with cerebral palsy. *Developmental medicine and child neurology*, 64(8), 1034–1043. <https://doi.org/10.1111/dmcn.15193>
- Moons KGM, Damen JAA, Kaul T, Hooft L, Andaur Navarro C, Dhiman P, Beam AL, Van Calster B, Celi LA, Denaxas S, Denniston AK, Ghassemi M, Heinze G, Kengne AP, Maier-Hein L, Liu X, Logullo P, McCradden MD, Liu N, Oakden-Rayner L, Singh K, Ting DS, Wynants L, Yang B, Reitsma JB, Riley RD, Collins GS, van Smeden M. PROBAST+AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. *BMJ*. 2025 Mar 24;388:e082505. doi: 10.1136/bmj-2024-082505. PMID: 40127903; PMCID: PMC11931409.
- Moons, K. G. M., Wolff, R. F., Riley, R. D., Whiting, P. F., Westwood, M., Collins, G. S., Reitsma, J. B., Kleijnen, J., & Mallett, S. (2019). PROBAST: A Tool to Assess Risk of Bias and Applicability of Prediction Model Studies: Explanation and Elaboration. *Annals of internal medicine*, 170(1), W1–W33. <https://doi.org/10.7326/M18-1377>
- Mueller, K. B., Hou, Y., Beach, K., & Griffin, L. P. (2024). Development and validation of a point-of-care clinical risk score to predict surgical site complication-associated readmissions following open spine surgery. *Journal of spine surgery (Hong Kong)*, 10(1), 40–54. <https://doi.org/10.21037/jss-23-89>
- Ouzzani, M., Hammady, H., Fedorowicz, Z. *et al.* Rayyan—a web and mobile app for systematic reviews. *Syst Rev* 5, 210 (2016). <https://doi.org/10.1186/s13643-016-0384-4>
- Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., McGuinness, L. A., ... Moher, D. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ (Clinical research ed.)*, 372, n71. <https://doi.org/10.1136/bmj.n71>
- Paranjape, K., Schinkel, M., Nannan Panday, R., Car, J., & Nanayakkara, P. (2019). Introducing Artificial

- Intelligence Training in Medical Education. *JMIR medical education*, 5(2), e16048. <https://doi.org/10.2196/16048>
- Park, S. H., Park, J., Hwang, J. H., Hwang, S. K., Hamm, I. S., & Kang, K. C. (2022). Risk factors for postoperative deep infection after instrumented spinal fusion surgery. *Journal of Clinical Medicine*, 11(3), 778.
- Peduzzi, P., Concato, J., Kemper, E., Holford, T. R., & Feinstein, A. R. (1996). A simulation study of the number of events per variable in logistic regression analysis. *Journal of clinical epidemiology*, 49(12), 1373–1379. [https://doi.org/10.1016/s0895-4356\(96\)00236-3](https://doi.org/10.1016/s0895-4356(96)00236-3)
- Pei, H., Wang, H., Chen, M., Ma, L., Liu, G., & Ding, W. (2021). Surgical site infection after posterior lumbar interbody fusion and instrumentation in patients with lumbar degenerative disease. *International wound journal*, 18(5), 608–615. <https://doi.org/10.1111/iwj.13562>
- Peng, W., Liang, Y., Lu, T., Li, M., Li, D. S., Du, K. H., & Wu, J. H. (2019). Multivariate analysis of incision infection after posterior lumbar surgery in diabetic patients: A single-center retrospective analysis. *Medicine*, 98(23), e15935. <https://doi.org/10.1097/MD.00000000000015935>
- Qiu, H. Y., Liu, D. M., Sun, F. L., Lu, C. B., Dai, J. J., Yang, Y. P., Huang, X. Y., Lei, W., & Zhang, Y. (2024). Development and validation of a clinical nomogram prediction model for surgical site infection following lumbar disc herniation surgery. *Scientific reports*, 14(1), 26910. <https://doi.org/10.1038/s41598-024-76129-y>
- Ramspek, C. L., Jager, K. J., Dekker, F. W., Zoccali, C., & van Diepen, M. (2020). External validation of prognostic models: what, why, how, when and where?. *Clinical kidney journal*, 14(1), 49–58. <https://doi.org/10.1093/ckj/sfaa188>
- Riley, R. D., Snell, K. I., Ensor, J., Burke, D. L., Harrell, F. E., Jr, Moons, K. G., & Collins, G. S. (2019). Minimum sample size for developing a multivariable prediction model: PART II - binary and time-to-event outcomes. *Statistics in medicine*, 38(7), 1276–1296. <https://doi.org/10.1002/sim.7992>
- Shah, A. A., Devana, S. K., Lee, C., Bugarin, A., Lord, E. L., Shamie, A. N., Park, D. Y., van der Schaar, M., & SooHoo, N. F. (2021). Prediction of Major Complications and Readmission After Lumbar Spinal Fusion: A Machine Learning-Driven Approach. *World neurosurgery*, 152, e227–e234. <https://doi.org/10.1016/j.wneu.2021.05.080>
- Song, Y. X., Yang, X. D., Luo, Y. G., Ouyang, C. L., Yu, Y., Ma, Y. L., Li, H., Lou, J. S., Liu, Y. H., Chen, Y. Q., Cao, J. B., & Mi, W. D. (2023). Comparison of logistic regression and machine learning methods for predicting postoperative delirium in elderly patients: A retrospective study. *CNS neuroscience & therapeutics*, 29(1), 158–167. <https://doi.org/10.1111/cns.13991>
- Tkatschenko, D., Hansen, S., Koch, J., Ames, C., Fehlings, M. G., Berven, S., Sekhon, L., Shaffrey, C., Smith, J. S., Hart, R., Kim, H. J., Wang, J., Ha, Y., Kwan, K., Hai, Y., Valacco, M., Falavigna, A., Taboada, N., Guirroy, A., Emmerich, J., ... Onken, J. (2023). Prevention of Surgical Site Infections in Spine Surgery: An International Survey of Clinical Practices Among Expert Spine Surgeons. *Global spine journal*, 13(7), 2007–2015. <https://doi.org/10.1177/21925682211068414>
- Van Calster, B., McLernon, D. J., van Smeden, M., Wynants, L., & Steyerberg, E. W. (2019). Calibration: the Achilles heel of predictive analytics. *BMC Medicine*, 17(1), 230. <https://doi.org/10.1186/s12916-019-1466-7>
- van Smeden, M., Reitsma, J. B., Riley, R. D., Collins, G. S., & Moons, K. G. (2021). Clinical prediction models: diagnosis versus prognosis. *Journal of clinical epidemiology*, 132, 142–145. <https://doi.org/10.1016/j.jclinepi.2021.01.009>
- Wang, H., Fan, T., Yang, B., Lin, Q., Li, W., & Yang, M. (2021). Development and Internal Validation of Supervised Machine Learning Algorithms for Predicting the Risk of Surgical Site Infection Following

- Minimally Invasive Transforaminal Lumbar Interbody Fusion. *Frontiers in medicine*, 8, 771608. <https://doi.org/10.3389/fmed.2021.771608>
- World Health Organization. (2022). Global Health Observatory: Obesity among adults, BMI ≥ 30 , prevalence (age-standardized estimate). Geneva: World Health Organization. Available at: [https://www.who.int/data/gho/data/indicators/indicator-details/GHO/prevalence-of-obesity-among-adults-bmi-30-\(age-standardized-estimate\)-\(-\)](https://www.who.int/data/gho/data/indicators/indicator-details/GHO/prevalence-of-obesity-among-adults-bmi-30-(age-standardized-estimate)-(-))
- Xiong, C., Zhao, R., Xu, J., Liang, H., Zhang, C., Zhao, Z., Huang, T., & Luo, X. (2022). Construct and Validate a Predictive Model for Surgical Site Infection after Posterior Lumbar Interbody Fusion Based on Machine Learning Algorithm. *Computational and mathematical methods in medicine*, 2022, 2697841. <https://doi.org/10.1155/2022/2697841>
- Zhang, Q., Chen, G., Zhu, Q., Liu, Z., Li, Y., Li, R., Zhao, T., Liu, X., Zhu, Y., Zhang, Z., & Li, H. (2024). Construct validation of machine learning for accurately predicting the risk of postoperative surgical site infection following spine surgery. *The Journal of hospital infection*, 146, 232–241. <https://doi.org/10.1016/j.jhin.2023.09.02>



10. Anexo

Anexo 1: Formulário De Seleção

Formulário De Seleção (screening)

Revisor A []

Revisor B []

1. Identificação do artigo

a) Sobrenome do autor:

País:

b) Nome do jornal:

c) Ano de publicação:

Volume:

2. Critérios de elegibilidadea) Inclui pacientes adultos (≥ 18 anos) ?

SIM []

NÃO []

b) Inclui pacientes submetidos a cirurgias de coluna?

SIM []

NÃO []

c) Inclui cirurgias eletivas ou relacionadas a condições degenerativas, traumáticas ou deformidades?

SIM []

NÃO []

d) O estudo avalia modelos preditivos de risco para infecções do local cirúrgico ou identifica fatores de risco pós-cirúrgicos?

SIM []

NÃO []

e) É um estudo empírico (RCT, coorte, transversal ou qualitativo)?

SIM []

NÃO []

f) Artigo em inglês ou português?

SIM []

NÃO []

Observações sobre o estudo:

3) Confirmação da elegibilidade/ Decisão final

Considerando os critérios de inclusão e exclusão e os objetivos da Revisão Sistemática, o estudo pode ser incluído?

[] INCLUIR

[] EXCLUIR

[] DÚVIDA (PRECISA DE CONSENSO OU SEGUNDO REVISOR)

Anexo 2: Formulário De Elegibilidade

Formulário De Elegibilidade (leitura completa do artigo)

Revisor A

Revisor B

1. Identificação do artigo

a) Sobrenome do autor:

País:

b) Nome do jornal:

c) Ano de publicação:

Volume:

2. Critérios de elegibilidade**a)** Inclui pacientes adultos (≥ 18 anos) submetidos a cirurgias de coluna lombar?SIM ☐NÃO ☐**b)** Avalia modelos preditivos de risco para infecções do local cirúrgico?SIM ☐NÃO ☐**c)** Modelos incluem fatores relacionados ao paciente, procedimento ou ambiente?SIM ☐NÃO ☐**d)** O estudo relata métricas de desempenho do modelo (sensibilidade, especificidade, precisão, AUC)?SIM ☐NÃO ☐**e)** Exclui pacientes com infecções pré-existentes?SIM ☐NÃO ☐**f)** Exclui cirurgias de emergência (e.g., trauma agudo)?SIM ☐NÃO ☐

Observações sobre o estudo:

3) Confirmação da elegibilidade/ Decisão final

Considerando os critérios de inclusão e exclusão e os objetivos da Revisão Sistemática, o estudo pode ser incluído?

☐ INCLUIR☐ EXCLUIR☐ DÚVIDA (PRECISA DE CONSENSO OU SEGUNDO REVISOR)