U.PORTO

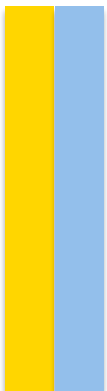
MESTRADO INTEGRADO EM MEDICINA
ANESTESIOLOGIA

"O Papel da Inteligência Artificial na Avaliação Pré-Anestésica do Risco Perioperatório"

Mariana da Naia Costa

M

2026



O Papel da Inteligência Artificial na Avaliação Pré-Anestésica do Risco Perioperatório

- REVISÃO NARRATIVA –

Dissertação de candidatura ao grau de Mestre em Medicina submetida ao Instituto de Ciências Biomédicas de Abel Salazar, Universidade do Porto

Estudante: Mariana da Naia Costa

Número de estudante: 201906983

Correio eletrónico: up201906983@up.pt

Mestrado Integrado em Medicina

Instituto de Ciências Biomédicas Abel Salazar, Universidade do Porto, Porto, Portugal

Orientador: Prof. Dr. Humberto José da Silva Machado

Assistente Hospitalar Graduado Sénior de Anestesiologia - Unidade Local de Saúde de Santo António, Porto, Portugal

Diretor do Serviço de Anestesiologia - Unidade Local de Saúde de Santo António, Porto, Portugal

Professor catedrático convidado do Mestrado Integrado em Medicina no Instituto Ciências Biomédicas Abel Salazar, Universidade do Porto, Porto, Portugal

Coorientadora: Dra. Débora Cavalheiro da Costa Domingues Marques

Assistente Hospitalar – Unidade Local de Saúde de Santo António, Porto, Portugal

Interna de Formação Específica de Anestesiologia

Junho, 2026

O Papel da Inteligência Artificial na Avaliação Pré-Anestésica do Risco Perioperatório

- REVISÃO NARRATIVA –

Dissertação de candidatura ao grau de Mestre em Medicina submetida ao Instituto de Ciências Biomédicas de Abel Salazar, Universidade do Porto

Assinado por: Mariana da Naia Costa
Identificação: BI14315856
Data: 2026-06-07 às 20:08:44

Mariana Naia Costa

Assinado por: Humberto José da Silva Machado
Identificação: BI05670003
Data: 2026-06-07 às 15:34:15

ɔsé da Silva Machado

Orientador

Assinado por: **Débora Cavalheiro da Costa Domingues Marques**
Num. de Identificação: 14625310
Data: 2026.06.07 20:18:39 +0100

CHAVE MÓVEL

Dra. Débora Cavalheiro da Costa Domingues Marques

Coorientadora

Junho, 2026

Declaração de integridade científica

Eu, Mariana da Naia Costa, declaro solenemente que esta dissertação, intitulada " O Papel da Inteligência Artificial na Avaliação Pré-Anestésica do Risco Perioperatório", é um trabalho original e genuíno. Durante o seu processo de investigação e elaboração, segui estritamente os princípios éticos e as normas académicas estabelecidas pela comunidade científica. Assumo total responsabilidade pelo conteúdo desta dissertação e estou ciente das possíveis consequências de qualquer violação dos princípios de integridade científica.

Assinado por: Mariana da Naia Costa
Identificação: BI14315856
Data: 2026-06-07 às 20:14:01

Mariana Naia Costa

Agradecimentos

Ao meu orientador Prof. Dr. Humberto Machado por ter aceitado orientar este trabalho e pelo compromisso com que acompanhou este percurso.
O estímulo ao pensamento crítico e à oportunidade de aprofundar o contacto com a anestesiologia constituíram contributos valiosos não só para o este trabalho como também para a minha formação médica.

À minha coorientadora Dra. Débora Cavalheiro da Costa Domingues Marques pelo apoio e sugestões no desenvolvimento deste trabalho.

À instituição que me acolheu ao longo destes anos de formação, por me ter proporcionado a oportunidade de concluir esta etapa e por ter contribuído para o meu crescimento académico, profissional e pessoal.

Por fim, um agradecimento especial à minha família e aos meus amigos pelo apoio incondicional, compreensão e motivação em todos os momentos.

Resumo

Contexto: A consulta pré-anestésica integra a avaliação do risco perioperatório individual, classicamente apoiada por *scores* clínicos (ASA-PS, RCRI, P-POSSUM, ARISCAT, EuroSCORE II, CFS) baseados em regressão logística e em combinações lineares de variáveis selecionadas. A informatização dos registos clínicos e o desenvolvimento de modelos de *Machine Learning* (ML) abriu a possibilidade de uma estratificação de risco multivariável, automatizada e potencialmente mais individualizada.

Objetivo: Sintetizar o estado da arte da aplicação das ferramentas de inteligência artificial (IA) na avaliação pré-anestésica e estratificação do risco perioperatório, descrever a evolução dos modelos preditivos, identificar metodologias de desenvolvimento e validação clínica, e discutir os obstáculos à sua implementação responsável na prática.

Métodos: Revisão narrativa de artigos publicados em *PubMed, Google Scholar, ScienceDirect, Scopus* e *Web of Science*, complementada por consulta em *ClinicalTrials.gov*. Foram considerados estudos originais, revisões sistemáticas, meta-análises e protocolos de ensaios clínicos sobre desenvolvimento, validação ou implementação de modelos de ML em adultos submetidos a cirurgia eletiva ou urgente inespecífica.

Resultados: Identificaram-se quatro gerações de modelos: Redes Neuronais Artificiais simples (ANN), modelos de *Ensemble* (*Random Forest, XGBoost, LightGBM)*, *Deep Neural Networks* e *Large Language Models*. Os modelos de *Ensemble* demonstram o melhor equilíbrio entre desempenho e transparência em conjuntos de dados de dimensão média (AUROC consistentemente entre 0.80 e 0.97), superando os *scores* tradicionais na predição de mortalidade, complicações cardiovasculares, lesão renal aguda e tempo de internamento. A maioria dos estudos é retrospetiva, com validação interna e elevado risco de viés segundo o PROBAST-AI. Os estudos com validação externa e prospetiva (Kotzé, Wingert, Rosen, Çatak, PEACH RCT, Periop ORACLE) e os ensaios em curso (IMAGINATIVE) começam a aproximar-se dos requisitos DECIDE-AI para implementação clínica adequada.

Conclusão: A IA na avaliação pré-anestésica progrediu da demonstração de conceito para aplicações clínicas com aprovação regulatória, mas a translação para a prática permanece ainda limitada pela escassez de validação externa, pelo viés algorítmico em populações sub-representadas, e por questões não consensuais de concordância, responsabilidade ética e legal. A adequação do algoritmo ao volume real de dados disponíveis, e não a complexidade do modelo por si só, destaca-se como determinante crítico da sua utilidade clínica.

PALAVRAS-CHAVE: Inteligência Artificial; *Machine Learning*; Anestesiologia; Avaliação Pré-anestésica; Registo Médico Eletrónico; Análise de Risco; Revisão Narrativa.

Abstract

Background: The preanesthetic evaluation is mainly characterized by its perioperative risk evaluation of each patient, according to pre-established clinical scores (ASA-PS, RCRI, P-POSSUM, ARISCAT, EuroSCORE II, CFS), which are based on logistic regression and linear combination of selected variables. The introduction of Electronic Health records in the digital era, combined with the development of Machine Learning Models allowed the possibility of multivariable risk stratification in a more automatic and potentially individualized way.

Then, the aim of this study is to synthetize the development of artificial intelligence to be applied in the preoperative period and risk assessment, understand its progress, identify the methodology frameworks, clinical validation and challenges against a responsible utilization of its tools on clinical practice.

Methods: The databases of *PubMed*, *Google Scholar*, *ScienceDirect, Scopus* e *Web of Science,* with addition of ClinicalTrials.gov for complementary information, were used for this narrative review to give the study a wide perspective of the matter. It was included original studies, systematic reviews, meta-analyses and clinical trials about the development, validation and implementation of Machine Learning models in adults submitted to unspecified elective or urgent surgery.

Results: Four generational models were identified: simple Artificial Neural Networks (ANN), Ensemble models (Random Forest, XGBoost, LightGBM), Deep Neural Networks and Large Language Models. The Ensemble Models revealed the best balance of optimal performance (AUROC values consistently between 0.80 and 0.97) with high interpretation, in median dimensional studies. Also, they stood out in predicting mortality, cardiovascular, ARI and hospitalization period comparatively to traditional scores. Most of the literature found were retrospective investigational studies with internal validation and high risk of bias, according to PROBAST-AI. The ones focused on external validation, prospective investigation (in Kotzé, Wingert, Rosen, Çatak, PEACH RCT, Periop ORACLE) and clinical trials (IMAGINATIVE) are still very few, but they highlight a better approach to DECIDE-AI requirements for an adequate clinical accomplishment.

Conclusion: The artificial intelligence's tools in pre-anesthetic evaluation evolved from its promising concept to clinical application with regulatory endorsement, except its practical usage remains limited by the lack of outer validation, algorithm bias in sub represented population, mixed professional positioning towards the use of the models and not consensual ethic and legal matters. Aligning the algorithm with the actual volume of available data, rather than relying on model complexity alone, emerges as a key determinant of its clinical value.

Keywords: Artificial Intelligence; Machine Learning; Anesthesiology; Preoperative Care; Electronic Health Records; Risk Assessment; Narrative Review

Índice

INTRODUÇÃO **1**

Contextualização **1**

Objetivos **2**

METODOLOGIA **3**

RESULTADOS **4**

Contextualização dos Modelos Preditivos **4**

Evolução e Desempenho dos Modelos *Machine Learning* **5**

Metodologias de Implementação e Validação Clínica **7**

Ferramentas de Avaliação Metodológica 7

Protocolos ERAS e Integração com ML 9

Validação Interna, Externa e Estudos Prospetivos 9

Estudos Prospetivos em Curso **10**

DISCUSSÃO **11**

Variabilidade do desempenho técnico dos modelos de IA **11**

Limitações metodológicas e viés **11**

Translação da predição para a clínica **12**

Validação externa e Estudos Prospetivos **13**

Desafios à implementação clínica — responsabilidade, XAI, equidade e ética **13**

CONCLUSÃO **15**

REFERÊNCIAS **16**

SIGLAS:

ACS NSQIP — American College of Surgeons National Surgical Quality Improvement Program

ANN — Artificial Neural Network (Rede Neuronal Artificial)

ARISCAT — Assess Respiratory Risk in Surgical Patients in Catalonia

ASA -PS — American Society of Anesthesiology Physical Status

AUC — Area Under the Curve (Área Sob a Curva)

AUROC — Area Under the Receiver Operating Characteristic Curve (Área Sob a Curva ROC)

CARES-ML — Combined Assessment of Risk Encountered in Surgery — Machine Learning

CE — Conformité Européenne (Marcação de Conformidade Europeia)

CFS — Clinical Frailty Scale (Escala de Fragilidade Clínica)

CONSORT-AI — Consolidated Standards of Reporting Trials — Artificial Intelligence

DECIDE-AI — Developing and evaluating Clinical AI systems Development and Evaluation

DL — Deep Learning (Aprendizagem Profunda)

DNN — Deep Neural Network (Rede Neuronal Profunda)

EHR — Electronic Health Record (Registo Médico Eletrónico)

ERAS — Enhanced Recovery After Surgery (Recuperação Acelerada Após Cirurgia)

EuroSCORE II — European System for Cardiac Operative Risk Evaluation II

FDA — Food and Drug Administration

GAM — Generalized Additive Model (Modelo Aditivo Generalizado)

GIST — Gastrointestinal Stromal Tumor (Tumor do Estroma Gastrointestinal)

GRADE — Grading of Recommendations Assessment, Development and Evaluation

HPI — Hypotension Prediction Index

HYPE — Hypotension Prediction RCT

IA — Inteligência Artificial

IC95% — Intervalo de Confiança a 95%

IMAGINATIVE — trial do CARES-ML no Singapore General Hospital

LIME — Local Interpretable Model-Agnostic Explanations

LLM — Large Language Model (Modelo de Linguagem de Grande Dimensão)

LRA — Lesão Renal Aguda

LightGBM- light gradient boosting Microsoft

MACCE — Major Adverse Cardiac and Cerebrovascular Events

MANOVA — Multivariate Analysis of Variance

ML — Machine Learning (Aprendizagem Automática)

NHS — National Health Service

NLP — Natural Language Processing (Processamento de Linguagem Natural)

NZRISK — score neozelandês de predição de mortalidade pós-operatória

ORACLE — Outcome Risk Assessment with Computer Learning Enhancement

OR — Odds Ratio

p — valor de probabilidade estatística

P-POSSUM — Portsmouth Physiological and Operative Severity Score for the enUmeration of Mortality and Morbidity

PEACH — PErioperative AI CHatbot

PLOS — Prolonged Length of Stay (Internamento Prolongado)

POSSUM — Physiological and Operative Severity Score for the enUmeration of Mortality and Morbidity

POSPOM — Preoperative Score to Predict Postoperative Mortality

PROBAST — Prediction model Risk of Bias ASsessment Tool

RCRI — Revised Cardiac Risk Index (Índice de Risco Cardíaco Revisto)

RCT — Randomized Controlled Trial (Ensaio Clínico Aleatorizado)

RF — Random Forest

RME — Registo Médico Eletrónico

ROC — Receiver Operating Characteristic

RR — Risk Ratio

SHAP — SHapley Additive exPlanations

SORT — Surgical Outcome Risk Tool

SUS — System Usability Scale

TAM — Technology Acceptance Model

TEP — Tromboembolismo Profundo

TVP — Trombose Venosa Profunda

UCI — Unidade de Cuidados Intensivos

XAI — Explainable Artificial Intelligence (Inteligência Artificial Explicável)

Introdução

Contextualização

A consulta pré-anestésica é fundamental para orientação do anestesista no período pré-operatório uma vez que integra a etapa fundamental de estratificação do risco perioperatório individual para otimização da gestão do doente cirúrgico. Permite orientar a decisão clínica sobre a estratégia anestésica e cuidados de segurança para reduzir a mortalidade e morbilidade associadas ao stress cirúrgico. Classicamente, esta avaliação de risco foi sempre apoiada por *scores* clínicos estabelecidos com base em métodos matemáticos probabilísticos de regressão logística simples para previsão de complicações. Contudo, estes modelos tradicionais, *scores* como o *Revised Cardiac Risk Index (*RCRI*)* e *Physiological and Operative Severity Score for the enUmeration of Mortality and Morbidity* (P-POSSUM*)*, estão limitados ao processamento de combinações lineares de variáveis e falham na clareza de interpretação e reconhecimento de padrões das mais complexas e na reprodutibilidade interindividual, juntamente com a variabilidade entre peritos na aplicação dos *scores* e de calibração da avaliação clínica [1].

Paralelamente ao desenvolvimento dos *scores* clínicos da avaliação pré-anestésica, emergiu também o conceito da inteligência artificial perioperatória através da introdução dos primeiros *expert systems* em codificar o conhecimento médico explícito para automatizar e agilizar o raciocínio clínico [2,3]. Posteriormente, foram integradas abordagens ontológicas, ou seja, de representação formal e estruturada da informação clínica para prevenção de erros e de situações de risco em tempo real [4], mas sempre dependente da codificação manual e incapazes de aprendizagem de processamento de dados clínicos autonomamente.

Deste modo, com a informatização dos registos clínicos e evolução dos programas informáticos aliados à necessidade de colmatar as limitações mantidas, começaram a surgir os modelos preditivos de inteligência artificial, *Machine Learning Models*, aplicados à medicina perioperatória e que assumem particular destaque na literatura atual [5–7]. Contrariamente às abordagens anteriores, a eficiência destas ferramentas digitais proporcionou o desenvolvimento de algoritmos complexos de multivariáveis, aprendizagem automática (*Machine Learning*) e identificação de fatores preditores de risco, potenciando uma utilidade promissora destes sistemas de apoio à decisão clínica em contexto de pré-operatório [8].

Assim, pretende-se que com uma rede de dados analíticos e estatísticos universal a partir de plataformas digitais, se consigam corrigir erros de variabilidade entre especialistas e de estratificação de risco. Neste contexto, a presente revisão narrativa procura sintetizar o estado da arte de utilização das ferramentas de IA na otimização da avaliação personalizada do risco perioperatório e melhor planeamento anestésico, sem descurar os desafios que se mantém à sua implementação, como a validação dos modelos preditivos em variadas situações clínicas, e questões de responsabilidade ética, para se considerar a extensão da sua utilização na prática.

Objetivos

Este estudo pretende contextualizar a aplicabilidade e utilidade clínica das inovações tecnológicas por meio de ferramentas de inteligência artificial, em contexto da fase pré-operatória e na consulta pré-anestésica, para fins de análise e estratificação do risco perioperatório individual.

Para o efeito, procura-se:

1. Enquadrar a evolução dos algoritmos de *Machine Learning,* no contexto histórico dos modelos preditivos em medicina perioperatória;
2. Identificar os modelos desenvolvidos para estratificação de risco e predição de *outcomes* perioperatórios, em contexto de avaliação pré-anestésica;
3. Analisar o desempenho dos modelos na discriminação de variáveis de risco, comparativamente aos *scores* clínicos tradicionais respetivos;
4. Apresentar as metodologias de desenho e validação dos modelos, assim como os estudos prospetivos para avaliação do outcome e capacidade de integração real na prática;
5. Destacar os principais obstáculos à implementação clínica, segurança e responsabilidade ética associada à utilização das ferramentas de IA.

Metodologia

Atendendo ao que se pretende investigar, optou-se por uma exposição narrativa que teve por base a pesquisa e revisão de artigos científicos obtidos através das bases de dados *PubMed*, *Google Scholar*, *Science Direct, Scopus*, *Web of Science* e consulta adicional de registo de ensaios clínicos em *ClinicalTrials.gov*, sem especificação da data de início e até ao mais recente publicado.

A revisão da literatura foi orientada de acordo com 3 pilares: o recurso à IA na avaliação pré-anestésica, a previsão de complicações no período perioperatório e a aplicabilidade na decisão clínica para gestão do pré, intra e pós-operatório.

Utilizaram-se as seguintes palavras-chave para pesquisa avançada:

- (“Perioperative Medicine” OR “Perioperative Period” OR “Perioperative Care”) AND “Risk Assessment” AND “Artificial Intelligence”;
- "Machine Learning" AND "Preoperative Assessment";
- "Machine Learning" AND "Anesthesia" AND "Risk Assessment"
- "Deep Learning" AND "Perioperative"
- "Natural Language Processing" AND "Anesthesia"
- "Large Language Models" AND "Preoperative"
- Combinações com "External Validation", "Electronic Health Records", "Postoperative Complications"

Como critérios de inclusão, foram considerados projetos originais, estudos de revisão, meta-análises e protocolos de ensaios clínicos que abordassem o desenvolvimento, validação ou implementação de modelos de Inteligência Artificial com dados pré-operatórios aplicado à avaliação pré-anestésica e à medicina perioperatória com foco na avaliação do risco perioperatório de doentes submetidos a cirurgia eletiva ou urgente, sem restrição quanto ao tipo de cirurgia e especialidade. Para contextualizar a discussão, foram incluídas referências adicionais identificadas durante a pesquisa, além das listas iniciais.

Excluíram-se os artigos exclusivamente sobre: cirurgia robótica sem componente de modelos preditivos de IA; sistemas de monitorização intraoperatória sem avaliação pré-operatória; artigos sobre oncologia fora do contexto perioperatório; estudos com populações muito específicas em amostra reduzida; imagiologia diagnóstica sem aplicação na medicina perioperatória; gestão da dor crónica sem relação com o perioperatório; estudos realizados no âmbito da medicina veterinária.

Os artigos identificados foram organizados em dois grupos temáticos de forma a facilitar a análise comparativa: o primeiro grupo agregou estudos focados no desenvolvimento técnico e validação de modelos preditivos, categorizados por tipo de algoritmo (modelos lineares clássicos, *Ensemble* paralelo ou sequencial, *deep learning* e modelos de linguagem) e *outcomes*; o segundo reuniu os estudos sobre estruturas metodológicas de avaliação, validação externa, implementação clínica e considerações éticas. Os artigos que abordavam múltiplos domínios foram classificados de acordo com o seu foco principal.

Resultados

Face aos objetivos que foram estabelecidos e assumidos como mote para o desenrolar desta narrativa, a literatura que a sustenta revelou que a maioria dos artigos encontrados sobre a aplicação de modelos de *Machine Learning* na medicina pré-operatória tratam-se de estudos de investigação em coorte retrospetivo sobre a capacidade de identificação de variáveis preditoras de risco com os algoritmos de IA em determinados domínios cirúrgicos, nomeadamente os que motivam uma investigação mais célere devido a complicações cirúrgicas importantes relatadas e o risco de mortalidade inerente, em comparação ao desempenho dos *scores* de risco tradicionais estabelecidos.

Mais recentemente, é possível evidenciar um reorientação no desenvolvimento da AI com o intuito da sua aplicação na prática clínica.

A partir de 2018, a literatura revista evidencia um crescimento exponencial da complexidade do processamento de dados clínicos do RME (Registo Médico Eletrónico) com os modelos de AI e em maior escala[9]. Além disso, verifica-se uma maior exigência da estruturação metodológica dos modelos (p.e. diretrizes de DECIDE-AI [10]), de protocolos estabelecidos para análise do seu desempenho (p.e. PROBAST adaptado[11]) e ferramentas para otimização da integração na prática hospitalar, através da transformação dos *outputs* computacionais em informação interpretável (p.e. com análise SHAP) e acionável para a colaboração e adesão dos clínicos[12,13]. O estudos prospetivos são escassos e generalistas na avaliação do risco perioperatório, não categorizando por tipo de cirurgia ou sistema orgânico os resultados do impacto dos modelos na prevenção de complicações.

Contextualização dos Modelos Preditivos

Os modelos *Machine Learning* são um subsistema da IA e integram a classe de modelos preditivos, mais precisamente simulam capacidades cognitivas humanas ao prever e realizar tarefas requeridas através de uma contínua aprendizagem de manipulação de múltiplos dados submetidos, reconhecimento autónomo de padrões e otimização de inferências estatísticas [14].

A compreensão da abordagem com os modelos ML requer o enquadramento dos modelos preditivos clássicos que sustentam as *guidelines* perioperatórias atuais na avaliação pré-anestésica. A maioria destes modelos têm por base a regressão logística e um número reduzido de variáveis selecionadas dos critérios clínicos pelos médicos para estratificação e previsão do risco perioperatório. Entre os mais referenciados nas *guidelines* atuais, o ASA *Physical Status (PS) score* permite estimar o risco anestésico global do indivíduo, mas não considera o plano cirúrgico e apresenta subjetividade importante; o RCRI e EuroSCORE II para risco cardiovascular em cirurgia não cardíaca e cardíaca, respetivamente, apesar da baixa reprodutibilidade dos resultados demonstrada; o ARISCAT para complicações pulmonares pós-operatórias, essencial para seleção e otimização pré-operatória, mas com fiabilidade limitada por viés externo e desempenho inconsistente para alguns grupos de doentes cirúrgicos; o P-POSSUM para risco de mortalidade abrangente em cirurgia major, que requer parâmetros de dados do período intraoperatório limitando a eficiência da predição na avaliação pré-anestésica; e o *Clinical Frailty Scale (CFS)* para estratificação da fragilidade perioperatória em doentes cirúrgicos idosos, apresenta um elevado índice de variabilidade entre peritos, introduzindo subjetividade e eficácia limitada[15]. Estes *scores* constituem o referencial de comparação utilizados pela maioria dos autores para avaliação do desempenho dos novos modelos computacionais.

Evolução e Desempenho dos Modelos *Machine Learning*

A ideia de utilização de modelos preditivos computacionais na medicina perioperatória surgiu em primeiro lugar com a integração dos modelos lineares de estatística clássicos, como é o caso da regressão logística, nas Redes Neuronais Artificiais (ANN) simples iniciais. Em particular, um dos estudos pioneiros para o estabelecimento deste conceito na gestão do período perioperatório permitiu evidenciar uma boa capacidade discriminativa de fatores de risco com modelos de ANN (AUROC = 0.9) e comparativamente com o método estatístico de multivariáveis (MANOVA) e a regressão logística simples, para previsão de necessidade de internamento prolongado numa amostra considerável de 960 doentes submetidos a cirurgia cardíaca e com recurso aos mesmos dados do pré-operatório e pós-operatório inicial utilizados nos outros dois modelos, contribuindo para redução do viés interno [16]. Embora haja uma clara falha na validação externa do modelo por falta de comparação com *scores* clínicos pertinentes estabelecidos previamente (EuroSCORE), a potencialidade dos modelos impulsionou a contínua exploração da sua capacidade em prever e reconhecer variáveis preditoras de risco na avaliação pré-operatória, à medida que se foram transformando em redes de análise mais complexa e estruturada.

Seguidamente, vieram os modelos de *Ensemble* de árvores que permitiram a integração de centenas de modelos baseados em regras (Árvores de Decisão), variáveis heterogéneas do Registo Médico Eletrónico e processamento de um volume de média dimensão de dados com um nível de compreensão inferior para os clínicos. O algoritmo *Random Forest* foi defendido pelos autores como o mais eficiente na estratificação de risco em múltiplos estudos superando os *scores* ASA-PS, POSPOM, Charlson [17,18], e ACS-NSQIP [19] na previsão de mortalidade pós-operatória; modelos RL e ANN na previsão de hipotensão pós-indução anestésica [20]; ACS-NSQIP [19] em eventos cardíacos e cerebrovasculares major (MACCE); e a decisão do clínico na previsão de múltiplas complicações simultâneas em cirurgia major, nomeadamente pulmonares, neurológicas e cardiovasculares, através da sua integração no modelo clínico *MySurgeryRisk* [21]. Adicionalmente, demonstrou um papel importante na predição de complicações em contextos cirúrgicos variados como fenómenos tromboembólicos e infeção pulmonar pós-operatórios [22] e hemorragia intraoperatória em neurocirurgia [23]. A título de curiosidade, apenas em contexto de uma amostra reduzida de doentes cirúrgicos com GISTs (*Gastrointestinal Stromal Tumor*), a regressão logística revelou superioridade clínica na discriminação e calibração da estratificação de risco de malignidade [24], destacando a dimensão amostral como possível determinante na adequação do algoritmo utilizado [25].

Uma variante metodológica de maior complexidade emergiu com os modelos de *Ensemble* sequencial, com maior ênfase nos algoritmos de *XGBoost* e *LightGBM*, e passaram a dominar a literatura revista a partir de 2020. Permitiram a construção sequencial de árvores de decisão e consequentemente a possibilidade de correção de erros das árvores anteriores com a aprendizagem automática, conferindo uma capacidade preditiva com maior acuidade em conjuntos de dados de grande dimensão com vantagem evidente sobre o *Random Forest* [19,25]. Esta superioridade foi demonstrada em contextos perioperatórios variados — especificamente na predição de mortalidade pós-operatória e de MACCE com o algoritmo *LightGBM*, num estudo prospetivo com coorte de mais de 1,4 milhões de doentes cirúrgicos em vinte instituições hospitalares distintas [19], na predição de tempo de internamento prolongado pós-operatório com *XGBoost* [26] e na predição de hipotensão pós-indução com *XGBoost*, atingindo AUROC superior a 0.85 [27]. Adicionalmente, a compatibilidade destes algoritmos com sistemas de Inteligência Artificial Explicável (XAI), permitiu solucionar a menor transparência dos modelos de *Ensemble*, através da explicação do raciocínio realizado pelo modelo

para atribuição de cada predição individual, transformando um valor estatístico numa fundamentação clínica estruturada. Esta família de técnicas que integra as ferramentas de análise SHAP, LIME e árvores interpretáveis permitiu aproximar os modelos dos requisitos de transparência e segurança necessários para a sua integração na consulta pré-anestésica tornando as predições individualmente interpretáveis [12,13]. Inclusivamente, a compreensão dos modelos traduziu-se em melhor integração na prática clínica [28], otimização e personalização do protocolo perioperatório ERAS (Enhanced Recovery After Surgery) por perfil de risco e doente [29,30].

Por sua vez, a evolução para arquiteturas de maior complexidade computacional surgiu com o subcampo de *Deep Learning* (DL). Definido como um campo computacional capaz de englobar modelos de múltiplas camadas neuronais com aprendizagem de representações hierárquicas progressivamente mais abstratas. Em contraste aos modelos de *Ensemble* que operam sobre variáveis clínicas estruturadas extraídas do Registo Médico Eletrónico, as redes neuronais profundas (DNN), modelos que constituem o DL, processam os dados através de camadas ocultas sucessivas, em que as iniciais identificam padrões simples e as mais profundas capturam interações complexas entre variáveis [14]. A principal limitação destas arquiteturas é a necessidade de grandes volumes de dados para treino adequado e a sua natureza de caixa negra, característica definida pela falta de transparência do raciocínio do sistema, que sem recurso a ferramentas de XAI, a contribuição de cada variável para a predição final é incompreensível aos clínicos.

A primeira aplicação dos DNN em contexto perioperatório correspondeu à predição de mortalidade pós-operatória, demonstrando inconsistência no desempenho entre dois estudos efetuados em 2019 e 2021 por *Fritz et al*.[31] (*British Journal of Anaesthesia)* e *Bonde et al*.[32] (*Lancet Digital Health)*, revelando que a complexidade do processamento de dados não se traduzia em capacidade preditiva superior aos modelos de *XGBoost*, em *outcomes* distintos. Este achado é metodologicamente relevante para a compreensão de que o volume e da natureza dos dados disponíveis são determinantes para a acuidade discriminativa destes modelos [25].

A expansão significativa das capacidades do *Deep Learning* ocorreu com o desenvolvimento do processamento de linguagem natural (NLP), a partir dos *Large Language Models* (LLM). O campo de NLP consiste na abordagem computacional capaz de processar e compreender a linguagem humana, permitindo a inclusão de dados clínicos não estruturados, nomeadamente notas clínicas, relatórios de anestesia e registos de consultas pré-anestésicas em texto livre. Vários estudos demonstraram um desempenho significativamente superior com a combinação dos modelos de DNN e NLP [33], essencialmente na predição de mortalidade pós-operatória face aos modelos baseados em dados estruturados RF e *XGBoost* e destacando o valor informativo que as narrativas clínicas pré-operatórias podem ter na acuidade preditiva e na concordância com a classificação ASA-PS manual dos anestesiologistas [34], com potencial para reduzir a variabilidade interobservador documentada na atribuição do ASA-PS score automatizado [1].

A fronteira mais recente do *Deep Learning* em medicina perioperatória é representada pelos *Large Language Models* (LLMs), que consistem em subarquiteturas treinadas em grandes volumes de texto que permitem raciocínio em linguagem natural sobre cenários clínicos complexos. O modelo de *GPT-4 Turbo* destacou-se no seu desempenho para previsão dos *outcomes* de mortalidade pós-operatória e admissão em Unidade de Cuidados Intensivos (UCI), contrariamente ao verificado para duração do internamento [35], revelando uma acuidade específica dependente da tarefa requisitada. No âmbito de complicações neurológicas e cardiovasculares, em 2025, um estudo retrospetivo unicêntrico revelou a superioridade de LLMs (BioGPT, ClinicalBERT e BioClinicalBERT) na predição dos *outcomes* LRA (Lesão Renal Aguda), TEP (tromboembolismo profundo), TVP (Trombose Venosa

Profunda), *delirium* e pneumonia (AUROC de 0,697 a 0,875), comparativamente a modelos tradicionais de NLP e ao *score* ACS-NSQIP, a partir de notas clínicas pré-operatórias provenientes de registos eletrónicos de anestesia [36].

Adicionalmente, os modelos de LLMs revelaram uma limitação específica atribuída ao seu risco de alucinação, isto é, criação de texto aparentemente confiante, mas factualmente incorreto, crítico para a fiabilidade das recomendações na estratificação de risco perioperatório [37].

Da literatura revista, destaca-se o estudo clínico randomizado com o modelo PEACH (*PErioperative AI CHatbot*) conduzido em Singapura [38], revelando o impacto clínico promissor da implementação de LLMs na medicina perioperatória, em contexto de consulta pré-anestésica. Trata-se do primeiro ensaio clínico aplicado exclusivamente a este momento do pré-operatório e com utilização específica do modelo *Claude 3.5 Sonnet*, distinguindo-se dos modelos preditivos anteriores por assistir no raciocínio clínico e não na categorização de um risco numérico, com integração de *guidelines* perioperatórias institucionais, a partir de 35 protocolos anestésicos, e dados dos doentes. O objetivo do estudo era perceber se o modelo através da sugestão de um perfil de risco individual, plano anestésico e referenciações necessárias, refletia uma melhoria da gestão de tempo da equipa clínica como outcome primário, mas apenas subgrupos analisados de doentes com complexidade moderada e peritos com maior grau de experiência verificaram uma redução e melhoria estatisticamente significativa do tempo de documentação e inclusão de lista de problemas. Na prática, a integração deste modelo traduz-se numa melhoria da capacidade de decisão clínica complementar ao trabalho do perito, além de um impacto económico considerável na instituição hospitalar em que se encontra. Na sequência destes resultados e do estudo retrospetivo prévio que demonstrou uma acuidade preditiva de elevada precisão, menos de 2% de alucinação e de 1% de desvio das *guidelines*, a *Health Sciences Authority* de Singapura aprovou o *PEACH* como dispositivo médico de *software* de Classe A, isto é, com um nível de menor risco aplicável a sistemas de apoio à decisão clínica, tendo sido implementado em uso clínico real na consulta pré-anestésica desde dezembro de 2024.

Resumidamente, a comparação do desempenho entre as quatro gerações de modelos revela uma progressão de complexidade que não se traduz linearmente em superioridade da sua capacidade preditiva. Os modelos de *Ensemble* continuam a demonstrar o melhor equilíbrio entre consistência de desempenho, nível de compreensão e em alcançar requisitos dos protocolos de avaliação de desempenho, com base em conjuntos de registos de dados de dimensão média, frequente nos hospitais europeus [25]. Por sua vez, as DNNs mostram vantagem em dados de alta dimensionalidade e não estruturados, mas não superam consistentemente o *XGBoost* [32]. Os LLMs acrescentam capacidade de processamento de notas clínicas livres e raciocínio clínico contextual, mas introduzem riscos de alucinação e requerem validação específica universal [35,37,38].

Metodologias de Implementação e Validação Clínica

Ferramentas de Avaliação Metodológica

O crescimento exponencial de modelos de ML perioperatórios na literatura motivou o desenvolvimento de *frameworks* metodológicos específicos para avaliar a sua qualidade, fiabilidade e adequação clínica. A avaliação do risco de viés dos modelos preditivos recorre ao PROBAST (*Prediction model Risk Of Bias ASsessment Tool*), uma ferramenta de quatro domínios que avalia a selecção dos participantes, a definição dos preditores, a especificação do *outcome* e a análise estatística (*Wolff et al*. 2019). A extensão formal da aplicação do PROBAST a modelos de ML perioperatórios adiciona

domínios de avaliação específica dos mesmos. Estes parâmetros consistem na fuga ou contaminação de dados (*data leakage)*, essencialmente caracterizada pela inclusão de variáveis adicionais do intra e pós-operatório e consequente inflação da discriminação do modelo, não reprodutível na prática; no *overfitting* ou sobreajuste, com aprendizagem de padrões particulares, ruído aleatório e tendências amostrais, perdendo a capacidade do modelo em generalizar o processamento correto para novos dados e fora da instituição de treino; e a seleção de hiperparâmetros, refletindo uma falsa taxa de aprendizagem, semelhante ao *data leakage,* e valores de AUROC sobrestimados. Na revisão sistemática de *Arina et al.* [11], a utilização do PROBAST-AI identificou que a maioria apresentava alto risco de viés externo, avaliação exclusiva pela métrica do AUROC sem calibração e sobreajuste não detetado, o que foi salvaguardado no estudo em *Mahajan et al*. [19] para estratificação do risco geral e cardiovascular perioperatório. Outros instrumentos quantitativos complementares ao AUROC identificados na literatura foram as métricas de sensibilidade e especificidade [40,34], o *Brier Score* para combinação de discriminação e calibração, mas dependente da prevalência do *outcome* [41,42] e o NRI e IDI para comparação do desempenho dos novos modelos aos anteriores [43], mas muito menos frequentemente expressos [44]. Esta conclusão tem implicações diretas para a interpretação dos resultados da literatura, sendo que um AUROC elevado não garante utilidade clínica real [40], a calibração, que representa correspondência entre probabilidades previstas e frequências observadas, é um requisito igualmente essencial para a aplicação segura e utilidade clínica [45].

Semelhante ao PROBAST-AI, ferramentas clássicas de *reporting*, como o TRIPOD em estudos de desenvolvimento e validação, o CONSORT-AI e o SPIRIT-AI para os ensaios clínicos, que serviram de base para protocolos de estudos prospetivos mais recentes IMAGINATIVE [46], Period ORACLE [47] e PEACH RCT [38].

A síntese da qualidade da evidência agregada é feita pelo GRADE (*Grading of Recommendations Assessment, Development and Evaluation*), embora não criado para avaliação de modelos preditivos ou IA, é um sistema universal de avaliação da força da evidência agregada disponível que classifica os estudos em quatro níveis, de muito baixo a alto. *Mehta et al*. [48], na revisão sistemática com meta-análise sobre intervenções aumentadas por ML nos cuidados perioperatórios, e *Shimada et al*. [49] classificaram a evidência maioritariamente nos níveis baixo a moderado, compatível com o predomínio dos estudos retrospetivos unicêntricos.

Em 2021, *Vasey et al*. [10] publicaram as diretrizes DECIDE-AI, o *framework* desenvolvido especificamente para *reporting* de estudos de implementação de IA clínica. Ao contrário do PROBAST que avalia o desenvolvimento do modelo — e do GRADE — que avalia a evidência agregada — o DECIDE-AI define requisitos para documentar a implementação em contexto real, abrangendo quatro domínios: o desempenho e segurança do modelo, utilidade prática e integração no sistema de trabalho da equipa, calibração e impacto clinicamente significativo. A análise da literatura identificada nesta revisão sugere que apenas um número muito limitado de estudos se aproxima dos requisitos DECIDE-AI — nomeadamente *Wingert et al*. [50], *Rosen et al*. [51] e o IMAGINATIVE *trial* em curso.

Por conseguinte, a avaliação da utilidade e implementação clínica das ferramentas de ML recorre a métodos específicos que complementam as métricas de desempenho técnico. A *System Usability Scale* (SUS) e o *Technology Acceptance Model* (TAM) são os instrumentos mais utilizados para medir, respetivamente, a utilidade percebida e os determinantes de aceitação tecnológica pelos clínicos, principalmente a compreensão da utilidade e a facilidade de utilização [52,53]. Métodos qualitativos como entrevistas estruturadas permitem identificar requisitos de utilidade não avaliados através de questionários, tendo sido utilizados no design da plataforma de visualização dos dados de ML por *Fritz et al*. [52], na avaliação das perspetivas dos doentes sobre IA na tomada de decisão

partilhada por *Gould et al*. [54] e no desenvolvimento de um *framework* de XAI centrado no utilizador, combinando ciclos iterativos com múltiplos grupos profissionais por *Davidson et al*. [55].

Protocolos ERAS e Integração com ML

O protocolo ERAS, desenvolvido inicialmente por Kehlet [56] para cirurgia colorrectal, é um conjunto de intervenções pré, intra e pós-operatórias destinadas a acelerar a recuperação cirúrgica. É Integrado desde a otimização nutricional pré-operatória e analgésica até à mobilização precoce, demonstrando uma redução consistente no tempo de internamento, complicações e custos hospitalares em múltiplas especialidades cirúrgicas [57]. A combinação entre o ERAS e os modelos de ML perioperatórios foi explorada por *Ripollés-Melchor et al.* [29], num estudo que utilizou análise SHAP para identificar os elementos do protocolo ERAS com maior impacto em doentes estratificados por nível de risco pré-operatório. Contudo, os resultados demonstraram que a adesão ao ERAS tem impacto diferenciado consoante o perfil de risco, ou seja, em doentes de risco intermédio a adesão tem o maior benefício marginal, enquanto em doentes de alto risco a contribuição não foi clinicamente significativa. Esta observação pode ter implicações para a consulta pré-anestésica, ao personalizar quais os elementos ERAS a priorizar em cada doente individualmente [29].

Validação Interna, Externa e Estudos Prospetivos

A progressão metodológica dos modelos de ML perioperatórios organiza-se em três níveis de evidência crescente. A primeira e mais expressa na literatura corresponde à investigação retrospetiva do desempenho dos modelos com validação interna, mais concretamente modelos treinados e testados em dados prévios do mesmo centro e fase temporal. Similarmente ao exposto em 2023, por Arina et al, a maioria dos estudos identificados nesta revisão enquadra-se nesta categoria.

No entanto, a validação externa está representada num número mais limitado nos estudos encontrados. Dos quais, em *Kotzé et al.* [40] desenvolveu modelos de ML para apoio à avaliação pré-operatória em dois NHS (*National Health Service) Trusts* britânicos com sistemas de EHR distintos, com coortes de desenvolvimento e validação externa de 110.732 e 67.878 doentes respetivamente. O modelo de predição do *score* ASA *Physical Status* obteve boa precisão e discriminação para doentes de menor risco (ASA-PS *score* 1 e 2), mas calibração inferior à validação interna. Já em *Vernooij et al.* [45], numa revisão sistemática de modelos de predição de mortalidade pós-operatória até 30 dias, reportou-se que foram avaliados quatro *scores*, entre eles o SORT e o P-POSSUM, com desempenho semelhante e estratificação de risco superior à observada. Exemplos adicionais de validação multicêntrica externa incluem o estudo de *Yoon et al.* [58], que validou externamente três coortes de hospitais distintos com um modelo de *Gradient Boosting* para predição de insuficiência respiratória após cirurgia não cardíaca, com validação da boa discriminação do modelo interna e externamente (AUROC > 0.82) a partir de 8 variáveis clínicas, mas sem comparação com o ARISCAT como em *Li et al.* [59]. O outro caso de estudo multicêntrico em *Li et al.* [60] integrou sinais vitais e indicadores laboratoriais extraídos do registo médico eletrónico e permitiu validar externamente a boa acuidade preditiva do modelo RF (AUROC de 0.923) para insuficiência cardíaca pós-operatória em 489 doentes submetidos a cirurgia não cardíaca, comparativamente a 7 outros algoritmos.

Os estudos prospetivos integram a minoria da literatura revista e consistem na avaliação do modelo aplicado nos doentes em tempo real, antes do conhecimento dos *outcomes*, reduzindo fenómenos como a contaminação de dados. Primeiramente, em *Wingert et al.* 50 reportou a validação prospetiva do estudo em 2019 (*Hill et al.* [17]) a partir da implementação em tempo real no pré e

intraoperatório de um modelo de *Random Forest* com 32 variáveis para previsão de risco geral incluindo mortalidade pós-operatória. Obteve-se uma discriminação elevada (AUROC 0.874), superior à classificação ASA-PS *score*, inferior ao modelo do estudo retrospetivo com RF e um maior número de variáveis, mas demonstrando viabilidade técnica e clínica. Seguidamente, em *Rosen et al.* [51] implementou um modelo de predição de mortalidade a 1 ano em cirurgia colorrectal oncológica numa coorte de 18.403 doentes, demonstrando redução de complicações médicas de cerca de 15% e de um terço das readmissões com o suporte do modelo ($p<0.001$), com AUROC de 0.79 na validação. Contudo, apenas no domínio exclusivo da consulta pré-anestésica, *Çatak et al.* [1] conduziram um estudo prospetivo que avaliou a concordância entre um sistema de IA e anestesiologistas para a classificação ASA-PS *score*, estratificando os resultados por género e anos de experiência dos clínicos. A concordância evidenciada entre anestesiologistas (*Krippendorff's alpha* ~0.74) foi boa e superior à verificada entre IA e estes, com concordância observada de aproximadamente 57%, refletindo a limitação do domínio subjetivo e experiência clínica na implementação destes sistemas.

No âmbito dos ensaios RCT, o Periop ORACLE [61] avaliou o impacto de predições de ML na tomada de decisão de clínicos em contexto perioperatório, demonstrando melhoria modesta na acuidade preditiva do risco e identificando o fenómeno de *automation bias*. O PEACH RCT [38], conduzido com 270 avaliações pré-operatórias no *Singapore General Hospital*, não demonstrou redução significativa no tempo global de documentação, mas demonstrou redução de tempo utilizado na análise de doentes de moderada complexidade (5.77 minutos, $p=0.010$) e em profissionais mais experientes (4.6 minutos, $p=0.040$), com preferência da documentação assistida pelo PEACH em mais de metade dos casos.

Estudos Prospetivos em Curso

Além dos estudos prospetivos publicados, foi possível evidenciar um conjunto de investigações em curso e outros com resultados por apurar.

Um com particular relevância é o ensaio IMAGINATIVE [46] com uma coorte de 9200 doentes, que se encontra a decorrer no *Singapore General Hospital* até 2027, cujo objetivo é avaliar o desempenho do modelo clínico CARES-ML, baseado no *Gradient Boosting*, na estratificação de risco pré-operatória, a partir da consulta pré-anestésica, e na melhoria da mortalidade pós-operatória e admissão em cuidados intensivos no decorrer do perioperatório, comparativamente aos cuidados standard.

Adicionalmente, um estudo prospetivo de coorte longitudinal realizado em duas instituições hospitalares em curso no Brasil[62] pretende desenvolver e validar uma ferramenta de IA baseada em LLM em português para avaliação pré-anestésica numa amostra de doentes adultos em cirurgia não cardíaca eletiva, com conclusão prevista para 2028, para averiguação da concordância da qualidade entre as ferramentas e peritos.

Três estudos prospetivos adicionais encontrados e concluídos não permitiram a extração de conclusões por falta de dados disponíveis à data. Nomeadamente os ensaios AI-PREOP[63], de 2025 no *Trabzon Kanuni Education and Research Hospital* na Turquia, que comparou o sistema de ML com a avaliação pré-operatória dos peritos; KIPeriOP, de 2024 no Hospital Universitário de *Würzburg*[64] na Alemanha, publicou resultados parciais na língua original sem publicação final internacional; e Aiinane, concluído em julho de 2025 no *Hospital Universitari de Bellvitge* em Barcelona[65], que avaliou prospetivamente um *software* de avaliação pré-anestésica com marcação CE Classe IIa.

Discussão

Variabilidade do desempenho técnico dos modelos de IA

Os resultados da presente revisão confirmam que os modelos de ML demonstram, na generalidade, uma superioridade evidente aos *scores* clínicos tradicionais, com AUROCs consistentemente entre 0.80 e 0.97, ou seja, com boa predição de mortalidade, complicações cardiovasculares, LRA e outros *outcomes* perioperatórios pertinentes, e comparativamente aos *scores* RCRI (0.68–0.75) e P-POSSUM nas mesmas amostras de doentes [44]. Esta superioridade é mais evidente com a crescente complexidade e dimensão dos conjuntos de dados, onde o algoritmo otimizado do modelo de *Ensemble* - o *XGBoost* - apresenta robustez notória [19,25].

Contudo, é também possível evidenciar alguma inconsistência no poder discriminativo demonstrado. Nomeadamente, os dois estudos conduzidos em 2022 e 2025 por *Kowadlo et al.* [66] e *Liang et al.* [25], respetivamente, revelam uma irregularidade possível no seu desempenho técnico ao mostrarem uma previsão sobreponível de mortalidade pós-operatória entre os modelos mais antigos e recentes, numa coorte de dimensão média de dados clínicos. No segundo caso, perante uma amostra mais pequena de doentes oncológicos gástricos, a regressão logística clássica destacou-se na discriminação e calibração da estratificação do risco de malignidade. Ambos tornam defensável o pressuposto de que a complexidade do modelo não traduz na prática uma previsão de risco com maior acuidade, antes as arquiteturas mais elaboradas algorítmicas devem ser proporcionais à dimensão de registos e prevalência do *outcome* em específico, consistente com o argumento em *Mann et al.* [25].

Esta consideração tem implicações chave para a prática clínica que toma lugar em contextos multicêntricos menores, como é o caso de Portugal, onde os conjuntos de dados perioperatórios podem não ter uma dimensão tão grande. Podendo neste intervalo de dados, algoritmos como o *Random Forest* e o *XGBoost*, representar escolhas metodológicas mais adequadas do que modelos de *deep learning* ou de processamento de linguagem natural, que requerem volumes de dados substancialmente maiores para um treino devido [25,14]. Desta maneira, o desvio do foco da literatura mais recente para os modelos de ML de alta complexidade não deve ser interpretado como evidência de que esses modelos são os mais precisos para implementação universal, sendo que o estudo e compreensão do algoritmo adaptado ao contexto real de dados disponível é um determinante crítico da sua qualidade preditiva.

Limitações metodológicas e viés

Um erro sistemático metodológico identificado ao longo dos estudos efetuados é a inconsistência no procedimento de avaliação comparativa do desempenho dos modelos de ML. Mais concretamente, a maioria dos investigadores compara os novos algoritmos com o ASA-PS ou a regressão logística genérica [18,20], independentemente das complicações avaliadas, não sendo específicos para o *outcome* ou cirurgia realizada. Ademais, na avaliação de hipotensão pós-indução, LRA e TVP pós-operatórias [20,67] não existem *scores* tradicionais capazes de validar o desempenho efetivo superior dos novos modelos, criando limitações na interpretação clínica sobre a relevância demonstrada [44,11]. Este viés metodológico não invalida os modelos desenvolvidos, mas seguramente que traz desconfiança na sua superioridade perante o estado da arte clínica.

O mesmo é argumentado pelos autores em *Arina et al.11* e *Vernooij et al.*[45], em 2023. Estes concluíram que, através do PROBAST, a maioria dos estudos mantém alto risco de viés no domínio da

análise estatística devido à avaliação quase exclusiva por métricas com o AUROC, negligenciando parâmetros de calibração, utilidade e impacto clínico na gestão do fluxo de trabalho, e validação externa escassa, com sobrestimação ou subestimação do risco absoluto em subgrupos [45,40]. Outro estudo exemplificativo, deste presente ano, ilustra esta dissociação ao identificar uma má generalização de resultados de modelos preditivos na integração clínica real quando comparados dois centros hospitalares com sistemas de registo médico eletrónico distintos, apontando para a necessidade de modelação local específica da predição de risco, uma vez que as probabilidades absolutas podem induzir uma decisão clínica desadequada, especialmente sobre intensidade de monitorização e otimização pré-operatória, porque nem sempre correspondem às frequências reais observadas[40].

Por fim, a predominância dos estudos retrospetivos em detrimento aos de validação externa prospetiva, traduzem pelo protocolo GRADE uma qualidade da evidência agregada baixa a moderada, o que reduz a confiança de recomendações clínicas a partir do corpo de estudos disponível [68,48,49].

Translação da predição para a clínica

A limitação identificada mais relevante da literatura, apesar da discussão entre as diferenças técnicas na discriminação e calibração, é a dificuldade da translação da investigação para a prática clínica, ou seja, a maioria dos autores defendem o desempenho analisado, mas não publicam sobre a segurança, utilidade e impacto na melhoria dos *outcomes* dos doentes a partir de integração na prática.

Dos poucos estudos encontrados, destacam-se o recente de *Wingert et al*. [50] - representa um marco na transição para aplicação prática segura ao demonstrar manutenção da acuidade do modelo de *Gradient Boosting* para predição de mortalidade pós-operatória em contexto real, ao mesmo que se aproxima das diretrizes DECIDE-AI*,* conferindo-lhe credibilidade. Contudo, sem avaliação do impacto clínico do modelo, nomeadamente melhoria de *outcome*s e adesão dos peritos.

Esta questão foi parcialmente abordada pelo *RCT Periop ORACLE* [61], que demonstrou apenas uma melhoria modesta na acuidade preditiva humana, ao fornecer previsões de risco de mortalidade pós-operatória através dos *ML* aos clínicos, introduzindo o fenómeno de *automation bias*, isto é, a tendência dos clínicos para adotarem as previsões do modelo em vez do próprio raciocínio clínico. Isto sugere que a disponibilização de um score de ML na consulta pré-anestésica é insuficiente, sendo necessários *frameworks* de decisão estruturada que orientem a interpretação e a resposta clínica aos *outputs* algorítmicos [69].

O único estudo pré-operatório com impacto clínico prospetivo demonstrado na presente revisão é o *Rosen et al.* [51], ainda que com uma coorte menor e específica da cirurgia geral. Este implementou um modelo de predição de mortalidade a 1 ano após cirurgia colorrectal oncológica, demonstrando uma redução estatisticamente significativa de cerca de 32% de incidência de complicações médicas ($p<0.001$) e uma redução de 35% das readmissões. Embora seja relevante salientar que o modelo apresentava uma AUC de apenas 0.79, inferior à maioria dos modelos da literatura perioperatória, ilustrando, assim, que a acuidade discriminativa elevada não é condição necessária ou suficiente para o impacto clínico significativo. O que determina o benefício é a integração do modelo em protocolos estruturados e acionáveis de cuidado perioperatório.

Outro estudo que se destacou pelo seu impacto clínico evidente foi o HYPE RCT, baseado no *Hypotension Prediction Index* (HPI, Edwards Lifesciences), desenvolvido em 2018 por *Hatib et al.* [70] em mais de 25.000 doentes. Constitui o único modelo ML perioperatório com aprovação regulatória da

FDA (*Food and Drug Administration*) e um impacto clínico efetivo demonstrado em múltiplos estudos. O HYPE RCT [71] demonstrou uma redução estatisticamente significativa na duração e profundidade da hipotensão intraoperatória, a partir de uma intervenção mais antecipada e dirigida, sem diferença significativa na dose cumulativa da fluidoterapia ou vasopressores administrados. A distinção entre o HPI e os modelos pré-operatórios, cuja tradução clínica permanece em grande medida por demonstrar [10], ilustra que a potencialidade dos modelos preditivos poderá estar limitada se apenas aplicada num registo de avaliação pré-anestésica, ou seja, a exploração da sua utilidade continuamente para a fase intraoperatória, poderá extrair uma maior acuidade na previsão e suporte da decisão clínica.

Validação externa e Estudos Prospetivos

A presente revisão identificou um número reduzido de estudos com validação externa efetiva. Em 2023, quantificou-se apenas uma pequena percentagem dos modelos perioperatórios identificados na revisão sistemática como externamente validados e em estudos multicêntricos, valor que reflete o padrão observado nos artigos desta revisão, onde a maioria dos estudos apresenta validação exclusivamente interna ou temporal. *Kotzé et al.* [40] e *Wingert et al.* [50] representam exceções notáveis, com validação geográfica externa e implementação em tempo real, respetivamente, constituindo marcos de referência para futuros estudos de implementação.

Para além dos escassos estudos publicados, importa reconhecer a existência de investigação prospetiva concluída cujos resultados permanecem por publicar (AI-PREOP 2025 Turquia[63], KIPeriOP 2024 Alemanha[64], *Aiinane* 2025 Barcelona[65])

O IMAGINATIVE *trial* [46], atualmente em curso em Singapura, é o maior RCT prospetivo em curso sobre ML na avaliação perioperatória com um bom alcance das metas traçadas pelos requisitos de DECIDE-AI. Avalia o *score* CARES-ML, na consulta pré-operatória, para melhores *outcomes* de mortalidade e admissão em UCI, e representa o nível de evidência de que se necessita para fundamentar as recomendações clínicas. Outro estudo prospetivo e de coorte longitudinal em dois centros hospitalares em curso até 2028, no Brasil[62], pretende desenvolver e validar uma ferramenta de IA baseada em LLM para avaliação pré-anestésica em cirurgia não cardíaca, permitindo evidenciar investigações em contextos linguísticos e geográficos atuais.

A existência de estudos concluídos sem publicação, traz-nos a possibilidade de que resultados negativos ou modestos permaneçam na literatura cinzenta, enquanto resultados positivos são publicados mais rapidamente, criando uma ilusão de perceção aumentada do benefício clínico destas ferramentas.

Desafios à implementação clínica — responsabilidade, XAI, equidade e ética

A implementação de ferramentas de ML na prática médica enfrenta desafios estruturais que refletem a dificuldade da sua integração na prática, indo além da questão da acuidade preditiva. O grau de interpretação dos modelos é um requisito de segurança reconhecido, especialmente em contextos de elevada criticidade que exigem compreensão dos mecanismos de lógica [13]. Efetivamente, com o aumento da complexidade e raciocínio abstrato dos algoritmos, especialmente de DL e LLMs, há um decréscimo da capacidade na sua interpretação. A explicabilidade de cada predição individual, realizada através da análise SHAP, é necessária à ação do anestesiologista para refutar ou aceitar de forma responsável, a recomendação do sistema de apoio [55,26,29,72,73]. Acrescenta-se que poderá ter precisamente uma implicação fundamental nas questões médico-legais de consentimento informado

e documentação explícita, como defendido por *Basile et al.* [74]. *Hofmann* [75] demonstrou que a forma como os *outputs* são apresentados clinicamente determina a sua utilidade efetiva, enfatizando a relevância da integração das dimensões técnica, clínica, ética e regulatória para que as ferramentas de ML na consulta pré-anestésica transitem da promessa investigacional para prática clínica segura e responsável. Por sua vez, como destacado pelas diretrizes DECIDE-AI [10], desenvolvidas precisamente para colmatar a lacuna entre o desenvolvimento técnico e a implementação clínica responsável, os requisitos de *reporting* que incluem segurança, utilidade, equidade e impacto no *workflow*, não se apresentam frequentemente nos estudos de desenvolvimento dos modelos identificados. A análise da literatura sugere que apenas um número muito reduzido de estudos, como o *Wingert et al*. [50], *Rosen et al. 5*[1] e o IMAGINATIVE *trial* (em curso), se aproxima dos requisitos DECIDE-AI.

As questões de falta de equidade algorítmica identificadas em alguns estudos representam implicações éticas a destacar. Por um lado, os modelos de ML são treinados em populações maioritariamente homogéneas e de etnia caucasiana, com alto risco de viés de seleção e externo [11], e mostram um desempenho tendencialmente inferior em grupos sub-representados e de rendimento médio-baixo nos dados de treino [76,69], o que pode perpetuar o impacto das desigualdades da qualidade no acesso ao cuidado diferenciado. Por outro lado, a uniformização dos cuidados não depende apenas das características da população amostral, tendo em conta que o perfil de experiência do clínico se revelou ser um definidor da concordância e postura perante a utilização de modelos de IA, levando a resultados discordantes [1].

Deste modo, é importante considerar que as questões de responsabilidade legal e ética na tomada de decisão assistida por IA continuam um domínio emergente e não consensual à data, também defendido em *The Lancet Regional Health Europe*[69], através do reconhecimento da necessidade de auditorias regulares.

Conclusão

Em jeito de conclusão, os novos modelos preditivos baseados em *Machine Learning* desafiam os *scores* tradicionais na acuidade de identificação de variáveis preditoras de risco e dos respetivos *outcomes* perioperatórios. A maioria das investigações no âmbito do desenvolvimento destes modelos para a avaliação pré-anestésica incide na estratificação de risco geral e na predição de mortalidade pós-operatória, sendo este o *outcome* mais consistentemente investigado na literatura.

Importa, contudo, sublinhar que a discriminação elevada demonstrada pelos modelos em múltiplas categorias de risco não se traduz necessariamente em calibração adequada, que é uma dissociação metodológica que altera a interpretação clínica imediata dos resultados reportados. A escolha dos algoritmos deve ser orientada pela dimensão amostral e natureza dos dados disponíveis, e não pela complexidade técnica do modelo, sendo que em conjuntos de dados de dimensão média, algoritmos como o *Random Forest* e o *XGBoost* demonstraram superioridade. Paralelamente, uma boa interpretação dos modelos é tão determinante quanto o seu desempenho preditivo para garantir uma utilização responsável e segura na consulta pré-anestésica.

As investigações mais recentes incidem sobre os LLMs, expandindo as capacidades do ML para o processamento de dados clínicos não estruturados. Contudo, o suporte de ensaios clínicos que demonstrem o seu impacto efetivo na prática perioperatória permanece insuficiente, mantendo esta área num estadio precoce de maturação clínica.

Em síntese, a presente revisão narrativa permite concluir que a inteligência artificial revela um papel promissor como ferramenta auxiliar da decisão clínica pré-anestésica, ainda que a sua plena integração esteja condicionada por desafios metodológicos, éticos e de implementação. Adicionalmente, apesar do tema central deste trabalho se centrar na consulta pré-anestésica, a literatura revista evidencia que uma proporção significativa dos estudos abrange fases distintas do perioperatório, sugerindo que o potencial da IA poderá desempenhar um papel igualmente significativo no intra e pós-operatório. Assim, as futuras investigações não devem deixar de contemplar todo o perioperatório para uma abordagem mais completa e precisa da gestão do risco do doente cirúrgico.

Referências

1. Çatak T, Aksu A, Saltalı AÖ, et al. Comparison of human–AI agreement in ASA scoring by gender and duration of clinical experience: a real-world study. BMC Med Inform Decis Mak. 2026; 26(1):66. doi:10.1186/s12911-026-03399-z
2. Duarte-Medrano G, Nuño-Lámbarri N, Paternò DS, et al. Advancing a Hybrid Decision-Making Model in Anesthesiology: Applications of Artificial Intelligence in the Perioperative Setting. Healthcare (Basel). 2026; 14(1):97. doi:10.3390/healthcare14010097
3. Dost A, Alaraj R, Mayet R, Agrawal DK. Reshaping Anesthesia with Artificial Intelligence: From Concept to Reality. Anesth Crit Care. 2025; 7(3):77-90. doi:10.26502/acc.087
4. Uciteli A, Neumann J, Tahar K et al. Ontology-based specification, identification and analysis of perioperative risks. J Biomed Semantics. 2017; 8(1):36. doi:10.1186/s13326-017-0147-8
5. Connor CW. Artificial Intelligence and Machine Learning in Anesthesiology. Anesthesiology. 2019; 131(6):1346-1359. doi:10.1097/ALN.0000000000002694
6. Hashimoto DA, Witkowski E, Gao L et al. Artificial Intelligence in Anesthesiology: Current Techniques, Clinical Applications, and Limitations. Anesthesiology. 2020; 132(2):379-394. doi:10.1097/ALN.0000000000002960
7. Paiste HJ, Godwin RC, Smith AD et al. Strengths-weaknesses-opportunities-threats analysis of artificial intelligence in anesthesiology and perioperative medicine. Front Digit Health. 2024; 6:1316931. doi:10.3389/fdgth.2024.1316931
8. Chinn NR, et al. Artificial intelligence-enabled clinical decision support systems in preadmission testing: a scoping review of risk prediction, triage, and perioperative workflows (2020–2025). [Detalhes de publicação a confirmar — disponível via ResearchGate, 2025]
9. Liu, K., Qiu, W., & Yang, X. Exploring the growth and impact of artificial intelligence in anesthesiology: a bibliometric study from 2004 to 2024. Frontiers in Medicine, 12. 2025; https://doi.org/10.3389/fmed.2025.1595060
10. Vasey B, Clifton DA, Collins GS, et al; DECIDE-AI Steering Group. DECIDE-AI: new reporting guidelines to bridge the development-to-implementation gap in clinical artificial intelligence. Nat Med. 2021; 27(2):186-187. doi:10.1038/s41591-021-01229-5
11. Arina P, Kaczorek MR, Hofmaenner DA, et al. Prediction of complications and prognostication in perioperative medicine: a systematic review and PROBAST assessment of machine learning tools. Anesthesiology. 2024;140(1):85-101. doi:10.1097/ALN.0000000000004764
12. Lundberg SM, Nair B, Vavilala MS, et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. Nat Biomed Eng. 2018;2(10):749-760. doi:10.1038/s41551-018-0304-0
13. Loftus TJ, Tighe PJ, Filiberto AC, et al. Artificial intelligence and surgical decision-making. JAMA Surg. 2019. doi:10.1001/jamasurg.2019.4917
14. Pettit RW, Fullem R, Cheng C, Amos CI. Artificial intelligence, machine learning, and deep learning for clinical outcome prediction. Emerg Top Life Sci. 2021;5(6):729-745. doi:10.1042/ETLS20210246
15. Bai, C., Xiao, F., Al-Ani et al. External Validation of an AI-based Preoperative Frailty Index using Real-World Data. The journals of gerontology. Series A, Biological sciences and medical sciences. 2025. https://doi.org/10.1093/gerona/glaf119

16. Rowan M, Ryan T, Hegarty F, O'Hare N. The use of artificial neural networks to stratify the length of stay of cardiac patients based on preoperative and initial postoperative factors. Artif Intell Med. 2007;40(3):211-221. doi:10.1016/j.artmed.2007.04.005
17. Hill BL, Brown R, Gabel E, et al. An automated machine learning-based model predicts postoperative mortality using readily-extractable preoperative electronic health record data. Br J Anaesth. 2019. doi:10.1016/j.bja.2019.07.030
18. Li Y, Wang J, Huang S, et al. Implementation of a machine learning application in preoperative risk assessment for hip repair surgery. BMC Anesthesiol. 2022;22. doi:10.1186/s12871-022-01648-y
19. Mahajan A, Esper S, Oo TH, et al. Development and validation of a machine learning model to identify patients before surgery at high risk for postoperative adverse events. JAMA Netw Open. 2023;6. doi:10.1001/jamanetworkopen.2023.22285
20. Kendale S, Kulkarni P, Rosenberg AD, Wang J. Supervised machine-learning predictive analytics for prediction of postinduction hypotension. Anesthesiology. 2018;129:675-688. doi:10.1097/ALN.0000000000002374
21. Bihorac A, Ozrazgat-Baslanti T, Ebadi A, et al. MySurgeryRisk: development and validation of a machine-learning risk algorithm for major complications and death after surgery. Ann Surg. 2019;269(4):652-662. doi:10.1097/SLA.0000000000002706
22. Xue B, Li D, Lu C, et al. Use of machine learning to develop and evaluate models using preoperative and intraoperative data to identify risks of postoperative complications. JAMA Netw Open. 2021;4. doi:10.1001/jamanetworkopen.2021.2240
23. Wang X, Li H, Fan X, Xuan Z, Xiao J, Xu W. Development of a preoperative prediction tool for massive intraoperative blood loss in spinal metastases surgery integrating MRI and clinical data: a multicenter study. Int J Surg. 2026 Feb 1;112(2):3975-3988. doi: 10.1097/JS9.0000000000003800. Epub 2025 Nov 4. PMID: 41186592.
24. Liang et al. (2025) — Journal of Gastrointestinal Surgery — DOI: 10.1016/j.gassur.2024.10.019 — PubMed / ScienceDirect
25. Mann J, Lyons MM, O'Rourke J, Davies SJ. Machine learning or traditional statistical methods for predictive modelling in perioperative medicine: a narrative review. J Clin Anesth. 2025;102:111782. doi:10.1016/j.jclinane.2025.111782
26. Cho H, Ahn I, Gwon H, et al. Explainable predictions of a machine learning model to forecast the postoperative length of stay for severe patients: machine learning model development and evaluation. BMC Med Inform Decis Mak. 2024;24. doi:10.1186/s12911-024-02755-1
27. Chen M, Zhang B, Liu Z, et al. The application of explainable artificial intelligence (XAI) in the prediction of intraoperative red blood cell transfusion in valvular heart surgery based on QLattice. BMC Med Inform Decis Mak. 2025;25. doi:10.1186/s12911-025-03081-w
28. El Hechi MW, Nour Eddine SA, Maurer LR, Kaafarani HMA. Leveraging interpretable machine learning algorithms to predict postoperative patient outcomes on mobile devices. Surgery. 2021;169(4):750–754. DOI: 10.1016/j.surg.2020.06.049
29. Ripollés-Melchor J, Aldecoa C, Sáez-Ruíz E, et al. Personalized risk prediction in colorectal surgery after Enhanced Recovery After Surgery protocol implementation: a machine learning approach. Anaesth Crit Care Pain Med. 2026;44(2):101535. doi:10.1016/j.accpm.2025.101535
30. Solanki SL, Pandrowala S, Nayak A, Bhandare M, Ambulkar RP, Shrikhande SV. Artificial intelligence in perioperative management of major gastrointestinal surgeries. World Journal of Gastroenterology. 2021;27(21):2758–2770. DOI: 10.3748/wjg.v27.i21.2758

31. Fritz BA, Cui Z, Zhang M, et al. Deep-learning model for predicting 30-day postoperative mortality. Br J Anaesth. 2019;123(5):688-695. doi:10.1016/j.bja.2019.07.025
32. Bonde A, Varadarajan KM, Bonde N, et al. Assessing the utility of deep neural networks in predicting postoperative surgical complications: a retrospective study. Lancet Digit Health. 2021;3(8):e471-e485. doi:10.1016/S2589-7500(21)00084-4
33. Chen PF, Chen L, Lin YK, et al. Predicting postoperative mortality with deep neural networks and natural language processing: model development and validation. JMIR Med Inform. 2022;10(5):e38241. doi:10.2196/38241
34. Chung J, Pyo H, Cho HJ, Park C. Performance of automated machine learning for assigning American Society of Anesthesiologists physical status classification in adult patients: a quantitative, retrospective analysis. BMC Anesthesiol. 2023;23(1):318. doi:10.1186/s12871-023-02248-0
35. Brodeur PG, Buckley TA, Kanjee Z, et al. Superhuman performance of a large language model on the reasoning tasks of a physician. arXiv. 2024. doi:10.48550/arxiv.2412.10849
36. Alba C, Xue B, Abraham J, Kannampallil T, Lu C. The foundational capabilities of large language models in predicting postoperative risks using clinical notes. NPJ Digit Med. 2025;8(1):95. doi:10.1038/s41746-025-01489-2
37. Komasawa N. Generative artificial intelligence revolution in safe anesthesia care: roles, transformations, and future perspectives. J Anesth. 2025;39(3):323-328. doi:10.1007/s00540-025-03468-z
38. Ke Y, Yang R, Lie SA, et al. Mitigating cognitive biases in clinical decision-making through multi-agent conversations using large language models: simulation study. NPJ Digit Med. 2025;8:399. doi:10.1038/s41746-025-01858-x
39. Moons KGM, Wolff RF, Riley RD, et al. PROBAST: a tool to assess risk of bias and applicability of prediction model studies: explanation and elaboration. Ann Intern Med. 2019;170(1):W1–W33.
40. Kotzé J, Brealey D, Singer M, Whittle J. Development, external validation and integration into clinical workflow of machine learning models to support pre-operative assessment in the UK. Anaesthesia. 2026. doi:10.1111/anae.16777
41. Assel M, Sjoberg DD, Vickers AJ. The Brier score does not evaluate the clinical utility of diagnostic tests or prediction models. Diagnostic and Prognostic Research. 2017;1:19. DOI: 10.1186/s41512-017-0020-3
42. Luo J, Lin S, Wang L, et al. Predicting Delayed Extubation After General Anesthesia in Postanesthesia Care Unit Patients Using Machine Learning: Model Development Study. JMIR Med Inform. 2025 Nov 11;13:e72602. doi: 10.2196/72602. PMID: 41218185; PMCID: PMC12604829.
43. Pencina MJ, D'Agostino RB Sr, D'Agostino RB Jr, Vasan RS. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. Statistics in Medicine. 2008;27(2):157–172. DOI: 10.1002/sim.2929
44. Bedford J, Fuest K, Maheshwari K. Implementing risk prediction models in perioperative care: a narrative review. Curr Opin Anaesthesiol. 2024;38(2):184-191. doi:10.1097/ACO.0000000000001445
45. Vernooij LM, van Klei WA, Moons KGM, et al. The comparative and added prognostic value of biomarkers to the Revised Cardiac Risk Index for preoperative prediction of major adverse

cardiac events and all-cause mortality in patients who undergo noncardiac surgery. Cochrane Database Syst Rev. 2021;12:CD013139. doi:10.1002/14651858.CD013139.pub2

46. Abdullah HR, Ke Y, Loh J, et al. IMpact of an Artificial intelligence-derived CARES risk score Generated by INtegrative Anaesthetic preoperaTIVE assessment (IMAGINATIVE) on improving perioperative outcomes: protocol for a randomised controlled superiority trial. BMJ Open. 2024;14(8):e081789. doi:10.1136/bmjopen-2023-081789
47. Fritz BA, King CR, Mehta D, Avidan MS, et al. Protocol for the Perioperative Outcome Risk Assessment with Cognitive Learning Engine (PERIOP ORACLE) randomized study. F1000Res. 2022;11:653. doi:10.12688/f1000research.122286.2
48. Mehta D, Gonnah AR, El-Helali R, Khan O. Machine learning-augmented interventions in perioperative care: a systematic review and meta-analysis. Br J Anaesth. 2024;133(6):1159-1172. doi:10.1016/j.bja.2024.08.007
49. Shimada T, Yokoyama Y, Murakami A, et al. Effect of artificial intelligence-assisted intraoperative hypotension prediction on postoperative outcomes: a systematic review and meta-analysis. BMC Anesthesiol. 2024;24:466. doi:10.1186/s12871-024-02853-7
50. Wingert T, Lee C, Pollard JB, et al. Prospective evaluation of a real-time deployed machine learning model to predict mortality and unanticipated intensive care unit admission after surgery. Br J Anaesth. 2026. doi:10.1016/j.bja.2025.11.042
51. Rosen JE, Birkmeyer JD, Berkowitz SA, et al. Adjuvant immunotherapy and 1-year mortality after resection of clinical stage I non-small cell lung cancer: a cluster-randomized trial of a machine learning-based decision aid. Nat Med. 2025. doi:10.1038/s41591-025-03942-x
52. Fritz BA, Abdelhack M, King CR, Chen Y, Avidan MS. Visualization and assessment of model selection for clinical predictive modeling. Anesth Analg. 2023;137(1):2-12. doi:10.1213/ANE.0000000000006577
53. Choudhury A. Factors influencing clinicians' willingness to use an AI-based clinical decision support system. Front Digit Health. 2022;4:920662. doi:10.3389/fdgth.2022.920662
54. Gould DJ, Glanville-Hearst M, Bunzli S, Choong PFM, Dowsey MM. Patients' views on AI for risk prediction in shared decision-making for knee replacement surgery: qualitative interview study. J Med Internet Res. 2023;25:e43632. doi:10.2196/43632
55. Davidson SR, Khairat S, Bunce A, et al. User-centred design framework for end-to-end clinical AI explainability. arXiv. 2025. doi:10.48550/arxiv.2504.02551
56. Kehlet H. Multimodal approach to control postoperative pathophysiology and rehabilitation. Br J Anaesth. 1997;78(5):606-617. doi:10.1093/bja/78.5.606
57. Gustafsson UO, Scott MJ, Hubner M, Nygren J, Demartines N, Francis N, et al. Guidelines for Perioperative Care in Elective Colorectal Surgery: Enhanced Recovery After Surgery (ERAS®) Society Recommendations: 2018. World Journal of Surgery. 2019;43(3):659–695. DOI: 10.1007/s00268-018-4844-y
58. Yoon SH, Han S, Jeon HS, et al. External validation of a machine learning model for postoperative respiratory failure prediction. Br J Anaesth. 2024. doi:10.1016/j.bja.2024.06.007
59. Li P, Chen Y, Xu W, et al. Machine learning models for predicting postoperative pulmonary complications: validation against the ARISCAT score. Br J Anaesth. 2024;132(5):904-913. doi:10.1016/j.bja.2024.02.025

60. Li Q, Liu Z, Yang Y, et al. Multicenter validation of a Random Forest model for predicting postoperative heart failure after noncardiac surgery. Front Med (Lausanne). 2025;12:1666885. doi:10.3389/fmed.2025.1666885
61. Fritz BA, King CR, Mehta D, et al; PERIOP ORACLE Research Group. Effect of machine learning-derived predictions on clinician assessment of postoperative complication risk: a randomised clinical trial. Br J Anaesth. 2024;133(5):1018-1027. doi:10.1016/j.bja.2024.08.004
62. ClinicalTrials.gov (Internet). Bethesda (MD): National Library of Medicine (US); NCT07290647: Prospective Validation of an Artificial Intelligence Tool for Pre-Anesthetic Assessment. Disponível em: https://clinicaltrials.gov/study/NCT07290647. Consultado pela última vez a 2026/05/16.
63. ClinicalTrials.gov (Internet). Bethesda (MD): National Library of Medicine (US); NCT07364942: Comparison of Artificial Intelligence and Anesthesiologist in Preoperative Risk Assessment (AI-PREOP). Disponível em: https://clinicaltrials.gov/study/NCT07364942. Consultado pela última vez a 2026/05/15.
64. ClinicalTrials.gov (Internet). Bethesda (MD): National Library of Medicine (US); NCT05284227: Artificial Intelligence-augmented Perioperative Clinical Decision Support (KIPeriOP). Disponível em: https://clinicaltrials.gov/study/NCT05284227. Consultado pela última vez a 2026/05/15.
65. ClinicalTrials.gov (Internet). Bethesda (MD): National Library of Medicine (US); NCT07021235: Proof of Concept of the Aiinane Preoperative Risk Assessment Tool (aiinane-24-1). Disponível em: https://clinicaltrials.gov/study/NCT07021235. Consultado pela última vez a 2026/05/15.
66. Kowadlo G, Mittelberg Y, Ghomlaghi M, et al. Development and validation of 'Patient Optimizer' (POP) algorithms for predicting surgical risk with machine learning. BMC Med Inform Decis Mak. 2024;24:70. doi:10.1186/s12911-024-02463-w
67. Ge X, Li M, Yin J, et al. Predicting acute kidney injury after noncardiac surgery using interpretable machine learning models. Sci Rep. 2025;15. doi:10.1038/s41598-025-05551-7
68. Bronnert R, Besch G, Hild O, et al. Performance of artificial intelligence models for predicting intraoperative complications during surgery in real time: a systematic review and meta-analysis protocol. BMJ Open. 2025;15:e106204. doi:10.1136/bmjopen-2025-106204.
69. Wilhelm TI, Rosen J, Vetter VM, et al. Artificial intelligence in clinical decision support systems for perioperative care: regulatory and ethical considerations. Lancet Reg Health Eur. 2024;46:101145. doi:10.1016/j.lanepe.2024.101145
70. Hatib F, Jian Z, Buddi S, et al. Machine-learning algorithm to predict hypotension based on high-fidelity arterial pressure waveform analysis. Anesthesiology. 2018;129(4):663-674. doi:10.1097/ALN.0000000000002300
71. Wijnberge M, Geerts BF, Hol L, et al. Effect of a machine learning-derived early warning system for intraoperative hypotension vs standard care on depth and duration of intraoperative hypotension during elective noncardiac surgery: the HYPE randomized clinical trial. JAMA. 2020;323(11):1052-1060. doi:10.1001/jama.2020.0592
72. Laios A, Kalampokis E, Johnson R, Munot S, Thangavelu A, Hutson R, et al. Factors Predicting Surgical Effort Using Explainable Artificial Intelligence in Advanced Stage Epithelial Ovarian Cancer. Cancers. 2022;14(14):3447. DOI: 10.3390/cancers14143447
73. Wang K, Lin L, Zheng R, Nan S, Lu X, Duan H. Leveraging large language models for preoperative prevention of cardiopulmonary bypass-associated acute kidney injury. Ren Fail. 2025;47(1):2509786. doi:10.1080/0886022X.2025.2509786.

74. Basile G, Petrucci G, Cascone L, et al. Artificial intelligence in anaesthesia: medico-legal implications of consent and accountability. Bioengineering (Basel). 2026;13(2):227. doi: 10.3390/bioengineering13020227
75. Hofmann B. The impact of output presentation format on clinical decision-making with artificial intelligence systems. Front Digit Health. 2026;8. doi:10.3389/fdgth.2026.[volume]
76. O'Reilly-Shah VN, Gentry KR, Walters AM, Zivot J, Anderson CT, Tighe PJ. Bias and ethical considerations in machine learning and the automation of perioperative risk assessment. Br J Anaesth. 2020;125(6):843-846. doi:10.1016/j.bja.2020.07.040

Bibliografia complementar

Brennan M, Puri S, Ozrazgat-Baslanti T, et al. Comparing clinical judgment with the MySurgeryRisk algorithm for preoperative risk assessment: a pilot usability study. Surgery. 2019;165(5):1035-1045.
Chiew CJ, Liu N, Wong TH, Sim YE, Abdullah HR. Utilizing machine learning methods for preoperative prediction of postsurgical mortality and intensive care unit admission. Ann Surg. 2020;272(6):1133-1139.
Giordano C, Brennan M, Mohamed B, Rashidi P, Modave F, Tighe P. Accessing artificial intelligence for clinical decision-making. Front Digit Health. 2021;3:645232.
Elfanagely O, Toyoda Y, Othman S, et al. Machine learning and surgical outcomes prediction: a systematic review. J Surg Res. 2021;264:346-361.
Abraham J, Bartek B, Meng A, et al. Integrating machine learning predictions for perioperative risk management: towards an empirical design of a flexible-standardized risk assessment tool. J Biomed Inform. 2022;137:104270.
Bellini V, Valente M, Bertorelli G, et al. Machine learning in perioperative medicine: a systematic review. J Anesth Analg Crit Care. 2022;2:2.
Kim S, Kwon S, Rudas Á, et al. Machine learning of physiologic waveforms and electronic health record data. Crit Care Clin. 2023;39(4):675-687.
Winter A, Pfitzner B, Van De Water R, et al. Overcoming the data barrier: transfer learning for 90-day mortality prediction in general surgery — a retrospective multicenter development and comparison study. Int J Surg. 2025;112:655-665.

INSTITUTO DE **CIÊNCIAS BIOMÉDICAS ABEL SALAZAR**