ReGeo ISSN: 2177-3246

ESTRATIFICAÇÃO DE RISCO CLÍNICO COM APRENDIZADO DE MÁQUINA: VALIDAÇÃO EXTERNA E CALIBRAÇÃO EM COORTES BRASILEIRAS

CLINICAL RISK STRATIFICATION USING MACHINE LEARNING: EXTERNAL VALIDATION AND CALIBRATION IN BRAZILIAN COHORTS

ESTRATIFICACIÓN DE RIESGO CLÍNICO CON APRENDIZAJE AUTOMÁTICO: VALIDACIÓN EXTERNA Y CALIBRACIÓN EN COHORTES BRASILEÑAS

https://doi.org/10.56238/revgeov17n10-119

Data de submissão: 04/09/2026 **Data de publicação:** 16/10/2026

Kesley Silveira de Albuquerque
Médico especialista em Clínica Médica
E-mail: kesleyalbuquerque@hotmail.com
Lattes: http://lattes.cnpq.br/7317979338995554

Jose Peixoto de Alencar Neto
Pós-graduado em Psiquiatria
E-mail: dr.alencarpeixoto77@gmail.com
Lattes: https://lattes.cnpq.br/0595996914155949

RESUMO

Modelos de estratificação de risco construídos com aprendizado de máquina só são úteis à decisão clínica se, além de ordenarem corretamente os pacientes, produzirem probabilidades fiéis e mantiverem desempenho fora da amostra em que foram treinados. As duas propriedades correspondentes, calibração e validação externa, foram examinadas neste trabalho em dois planos: o da norma que regula o software como dispositivo médico no Brasil e o da produção científica brasileira. Analisou-se integralmente a Resolução da Diretoria Colegiada nº 657, de 24 de março de 2022, e construiu-se um corpus de 859 estudos com afiliação brasileira sobre modelos preditivos clínicos com aprendizado de máquina, publicados entre 2020 e 2026 e indexados no Europe PMC. Dos 788 com resumo disponível, 296 relatam medida de discriminação; destes, obteve-se o texto integral de 222. A resolução, com 19.616 caracteres e 25 artigos, emprega a palavra validação 11 vezes e nenhuma vez os termos calibração, discriminação, viés, subgrupo, generalização ou validação externa. Entre os 222 artigos, a calibração é mencionada em 12,6% dos resumos e em 45,0% dos textos integrais, e a validação externa em 15,8% e 62,6%. Curva ou inclinação de calibração aparece em 21,6% dos textos e o guia TRIPOD em 18,5%. Um quarto dos artigos, 25,7%, não menciona nenhuma das duas propriedades.

Palavras-chave: Modelos Preditivos. Calibração. Validação Externa.

ABSTRACT

Risk stratification models built with machine learning are only useful for clinical decisions if, besides ranking patients correctly, they produce faithful probabilities and retain performance outside the sample in which they were trained. The two corresponding properties, calibration and external validation, were examined here in two planes: the rule governing software as a medical device in Brazil and Brazilian scientific output. Collegiate Board Resolution No. 657 of 24 March 2022 was analysed in full, and a corpus was built of 859 studies with Brazilian affiliation on clinical predictive models using machine learning, published between 2020 and 2026 and indexed in Europe PMC. Of the 788

with available abstracts, 296 report a discrimination measure; full text was obtained for 222 of these. The resolution, with 19,616 characters and 25 articles, uses the word validation 11 times and never uses the terms calibration, discrimination, bias, subgroup, generalisation or external validation. Among the 222 articles, calibration is mentioned in 12.6% of abstracts and 45.0% of full texts, and external validation in 15.8% and 62.6%. A calibration curve or slope appears in 21.6% of full texts and the TRIPOD guideline in 18.5%. A quarter of the articles, 25.7%, mention neither property.

Keywords: Prediction Models. Calibration. External Validation.

RESUMEN

Los modelos de estratificación de riesgo construidos con aprendizaje automático solo son útiles para la toma de decisiones clínicas si, además de ordenar correctamente a los pacientes, producen probabilidades fiables y mantienen desempeño fuera de la muestra en la que fueron entrenados. Las dos propiedades correspondientes, calibración y validación externa, fueron examinadas en este trabajo en dos planos: el de la norma que regula el software como dispositivo médico en Brasil y el de la producción científica brasileña. Se analizó íntegramente la Resolución de la Junta Directiva nº 657, de 24 de marzo de 2022, y se construyó un corpus de 859 estudios con afiliación brasileña sobre modelos predictivos clínicos con aprendizaje automático, publicados entre 2020 y 2026 e indexados en Europe PMC. De los 788 con resumen disponible, 296 reportan medida de discriminación; de estos, se obtuvo el texto completo de 222. La resolución, con 19.616 caracteres y 25 artículos, utiliza la palabra validación 11 veces y ninguna vez los términos calibración, discriminación, sesgo, subgrupo, generalización o validación externa. Entre los 222 artículos, la calibración se menciona en el 12,6% de los resúmenes y en el 45,0% de los textos completos, y la validación externa en el 15,8% y el 62,6%. Curva o pendiente de calibración aparece en el 21,6% de los textos y la guía TRIPOD en el 18,5%. Una cuarta parte de los artículos, el 25,7%, no menciona ninguna de las dos propiedades.

Palabras clave: Modelos Predictivos. Calibración. Validación Externa.

1 INTRODUÇÃO

Um modelo de estratificação de risco clínico cumpre duas funções distintas. A primeira é ordenar: separar quem tem maior probabilidade de um desfecho de quem tem menor. Essa capacidade é medida por índices de discriminação, entre os quais a área sob a curva característica de operação do receptor, universalmente relatada pela sigla AUC. A segunda função é quantificar: quando o modelo afirma que um paciente tem 20% de risco, é preciso que, entre cem pacientes aos quais ele atribua esse valor, cerca de vinte apresentem o desfecho. Essa propriedade é a calibração, e nenhum índice de discriminação a mede.

A distinção não é acadêmica. Dois modelos com a mesma AUC podem produzir probabilidades radicalmente diferentes, e é sobre a probabilidade, não sobre a ordenação, que se decide internar, operar ou encaminhar. Um modelo mal calibrado que superestime o risco gera intervenção desnecessária; um que o subestime retarda cuidado. A ordenação pode estar correta nos dois casos.

A segunda exigência é a validação externa. Um modelo ajustado em uma coorte aprende também as particularidades daquela população, daquele hospital e daquele período. Reproduzir o desempenho em dados independentes, de outro serviço, de outra região ou de outro momento, é o que distingue um modelo transportável de um modelo que descreve bem apenas a amostra de origem. Em um país de dimensões continentais e desigualdade assistencial pronunciada, essa transportabilidade não pode ser presumida.

No plano regulatório, o uso clínico desses modelos é alcançado pela Resolução da Diretoria Colegiada nº 657, de 24 de março de 2022, que dispõe sobre a regularização de software como dispositivo médico. A resolução entrou em vigor em 1º de julho de 2022 e, na data desta análise, contava 1.569 dias de vigência. Ela exige validação, mas convém verificar com que vocabulário técnico a exige.

Três perguntas orientam a análise. Primeiro, que propriedades de desempenho a norma brasileira efetivamente nomeia. Segundo, com que frequência a literatura brasileira sobre modelos preditivos clínicos relata calibração e validação externa, e em que parte do artigo. Terceiro, se há tendência de melhora ao longo do período examinado.

2 METODOLOGIA

A pesquisa combina análise documental de norma e análise bibliométrica de conteúdo. O componente normativo examina o texto integral da Resolução da Diretoria Colegiada nº 657, de 24 de março de 2022, publicada no Diário Oficial da União de 30 de março de 2022, com 19.616 caracteres e 25 artigos. Contaram-se as ocorrências de termos técnicos do campo dos modelos preditivos,

verificando-se para cada ausência o contexto das ocorrências próximas, a fim de evitar falsos negativos por variação de grafia.

O componente bibliométrico partiu de consulta ao Europe PMC, em 17 de outubro de 2026, com estratégia que exige simultaneamente afiliação brasileira, método de aprendizado de máquina, finalidade de modelo preditivo ou de estratificação de risco, publicação entre 2020 e 2026 e presença dos termos pacientes ou coorte, excluindo-se revisões e estudos veterinários identificados pelo título. A consulta retornou 859 registros, dos quais 788 com resumo disponível.

Sobre os 788 resumos aplicou-se detecção léxica de cinco marcadores: discriminação, por AUC, AUROC, área sob a curva, estatística C ou índice de concordância; calibração, por calibração, Brier, Hosmer-Lemeshow ou diagrama de confiabilidade; validação externa, por validação externa, coorte ou conjunto independente, validação temporal, geográfica ou multicêntrica; validação interna, por validação cruzada, divisão treino-teste, reamostragem ou k-partições; e menção ao TRIPOD, guia internacional de relato de modelos preditivos.

Dos 788 resumos, 296 relatam medida de discriminação, critério adotado para identificar os estudos que efetivamente desenvolvem ou avaliam um modelo. Desses, 249 possuem identificador no PubMed Central, e obteve-se o texto integral de 222, com extensão mediana de 36.217 caracteres após remoção de referências, tabelas e figuras. Sobre esses textos aplicou-se a mesma detecção léxica, acrescida de dois marcadores mais exigentes: curva, gráfico ou inclinação de calibração, e escore de Brier. Os percentuais de resumo e de texto integral referem-se sempre ao mesmo conjunto de 222 artigos, o que torna as duas colunas diretamente comparáveis.

A varredura de literatura de apoio priorizou periódicos brasileiros das séries acompanhadas por esta linha de trabalho, entre eles Aracê, Lumen et Virtus e títulos indexados na SciELO, além de periódicos de direito sanitário e saúde coletiva. Todas as referências foram verificadas pelo identificador digital de objeto junto ao registro do editor, conferindo-se autoria, periódico, volume, número, páginas e ano. Não se apresentam quantitativos de documentos triados, pois o levantamento não seguiu protocolo censitário.

Três cautelas delimitam o alcance do que segue. A primeira, e mais importante, é que a detecção léxica identifica a presença do termo, não a adequação do procedimento: um artigo que escreva a palavra calibração pode tê-la avaliado de modo insuficiente, e a contagem não distingue os dois casos. A segunda é que o corpus é o conjunto indexado no Europe PMC com afiliação brasileira, o que exclui periódicos nacionais não indexados nessa base e inclui estudos de coautoria internacional cuja coorte pode não ser brasileira. A terceira é que o recorte de 222 textos integrais privilegia artigos de acesso aberto, que podem diferir sistematicamente dos demais quanto a práticas de relato.

3 DESENVOLVIMENTO

3.1 O VOCABULÁRIO DA NORMA

A Resolução nº 657/2022 regula a regularização de software como dispositivo médico, categoria que abrange os sistemas de apoio à decisão clínica baseados em modelos preditivos. O seu § 4º determina que as evidências de validação devem ser suficientes para garantir exatidão, confiabilidade e desempenho pretendido consistentes, e outro dispositivo exige que o software seja concebido de modo a garantir a repetibilidade, a confiabilidade e o desempenho de acordo com a utilização prevista. O § 5º concedeu aos serviços de saúde prazo de dois anos, contados da publicação, para realizar a validação de software desenvolvido internamente, prazo esgotado em 30 de março de 2024.

A contagem de termos no texto integral revela o vocabulário escolhido. A palavra validação aparece 11 vezes, clínico 21, desempenho 7, atualização 7, confiabilidade 2, monitoramento 2, exatidão 1 e repetibilidade 1. Em contrapartida, os termos calibração, discriminação, viés, subgrupo e generalização registram, cada um, zero ocorrência. A palavra externa aparece uma única vez, e em acepção alheia ao tema: refere-se a fatores externos de ambiente, como nível de luz e de ruído, na especificação de plataformas móveis.

Quadro 1. Perfil lexical da Resolução da Diretoria Colegiada nº 657/2022

Termo	Ocorrências	Campo a que pertence	Situação na norma
validação	11	requisito geral de evidência	núcleo do texto
clínico	21	finalidade e evidência	núcleo do texto
desempenho	7	propriedade exigida	exigido sem especificação
confiabilidade	2	propriedade exigida	exigido sem especificação
exatidão	1	propriedade exigida	exigido sem especificação
repetibilidade	1	propriedade exigida	exigido sem especificação
calibração	0	qualidade da probabilidade	ausente
discriminação	0	capacidade de ordenar	ausente
validação externa	0	transportabilidade	ausente
viés	0	equidade e subgrupos	ausente
subgrupo	0	equidade e subgrupos	ausente
generalização	0	transportabilidade	ausente

Fonte: elaboração própria a partir do texto integral da resolução, com 19.616 caracteres e 25 artigos.

Dois outros elementos do texto merecem registro porque tocam no ponto em que a calibração se degrada. A resolução emprega a palavra atualização 7 vezes e monitoramento 2 vezes, e disciplina as modificações que criem novas funcionalidades ou indicações de uso clínicas, afetem

significativamente as funcionalidades clínicas, a segurança, a eficácia ou o desempenho. Um modelo preditivo em uso, porém, pode perder calibração sem que nenhuma modificação seja feita nele, apenas porque a população atendida, a prevalência do desfecho ou a prática assistencial mudaram. Esse deslocamento silencioso é exatamente o que um regime de monitoramento de desempenho precisaria captar, e é a propriedade que o texto não nomeia.

O § 5º do mesmo dispositivo concedeu aos serviços de saúde prazo de dois anos, contados da publicação em 30 de março de 2022, para realizar a validação de software desenvolvido internamente, e o § 3º prevê que a ausência dos registros de validação por pelo menos dez anos após o descarte torna o software não regularizado, sujeitando o serviço a penalidades sanitárias. O prazo esgotou-se em 30 de março de 2024, há 931 dias na data desta análise. A obrigação, portanto, é exigível há mais de dois anos e meio, e recai sobre serviços que desenvolvem os seus próprios modelos, categoria que inclui hospitais universitários e grandes redes assistenciais.

O Quadro 1 sintetiza o resultado. A norma exige desempenho, exatidão, confiabilidade e repetibilidade, que são propriedades de um sistema de software, e não nomeia as propriedades estatísticas pelas quais um modelo preditivo é avaliado na literatura metodológica. Isso não significa que a exigência de validação seja vazia, pois um conceito aberto pode ser preenchido por normas técnicas e pela prática de avaliação da agência. Significa que o texto não fixa, por si, o critério contra o qual a calibração ou a transportabilidade de um modelo seriam aferidas. A adequação do desenho regulatório brasileiro para software como dispositivo médico vem sendo discutida na literatura nacional, com atenção especial à equidade e à inclusão (Jambor, 2025; Morais, 2025).

3.2 O QUE A LITERATURA BRASILEIRA RELATA

A consulta ao Europe PMC retornou 859 estudos com afiliação brasileira sobre modelos preditivos clínicos com aprendizado de máquina publicados entre 2020 e 2026, dos quais 788 com resumo e 730 em acesso aberto. Entre os 788 resumos, 296, ou 37,6%, relatam alguma medida de discriminação. Esse subconjunto é o que interessa, pois reúne os estudos que de fato avaliam um modelo. Nele concentra-se a análise seguinte, com os 222 artigos cujo texto integral foi obtido.

Convém observar o que fica de fora desse recorte. Dos 788 resumos, 492, ou 62,4%, não relatam medida de discriminação. Parte expressiva desse grupo são estudos que empregam aprendizado de máquina para finalidades que não a predição de risco em pacientes, como classificação de imagens, análise de sinais ou agrupamento exploratório, e para os quais a AUC pode não ser a métrica pertinente. Outra parte são estudos de modelos preditivos cujo resumo simplesmente não traz a métrica. Como não é possível separar os dois casos por detecção léxica, optou-se pelo critério

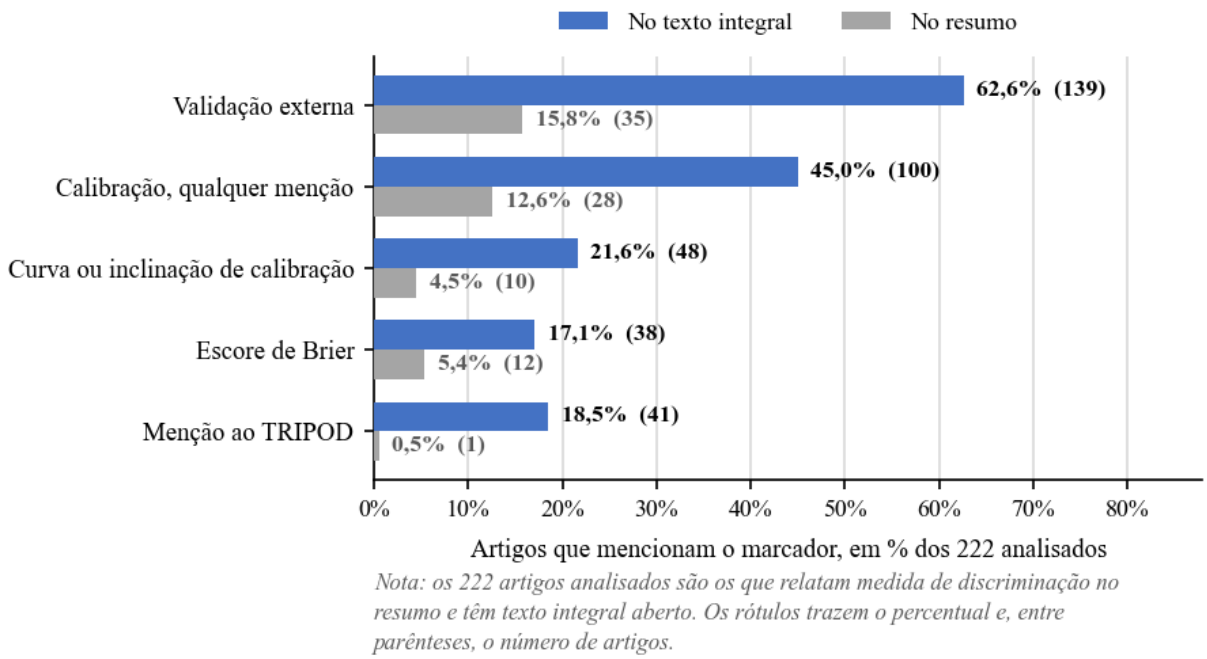
conservador de restringir a análise ao subconjunto que explicitamente relata discriminação, o que torna o denominador mais exigente e os percentuais mais defensáveis.

Quadro 2. Construção do corpus analisado

Etapa	Critério	Registros
Consulta ao Europe PMC	afiliação brasileira, aprendizado de máquina, modelo preditivo, 2020 a 2026, sem revisões	859
Com resumo disponível	campo de resumo preenchido	788
Com medida de discriminação no resumo	AUC, AUROC, área sob a curva, estatística C ou concordância	296
Com identificador no PubMed Central	condição para obtenção do texto integral	249
Com texto integral obtido e analisado	arquivo válido com mais de 4.000 caracteres úteis	222

Fonte: elaboração própria a partir de consulta ao Europe PMC em 17 de outubro de 2026.

Gráfico 1. Marcadores metodológicos no resumo e no texto integral dos mesmos 222 artigos



Nota: os 222 artigos analisados são os que relatam medida de discriminação no resumo e têm texto integral aberto. Os rótulos trazem o percentual e, entre parênteses, o número de artigos.

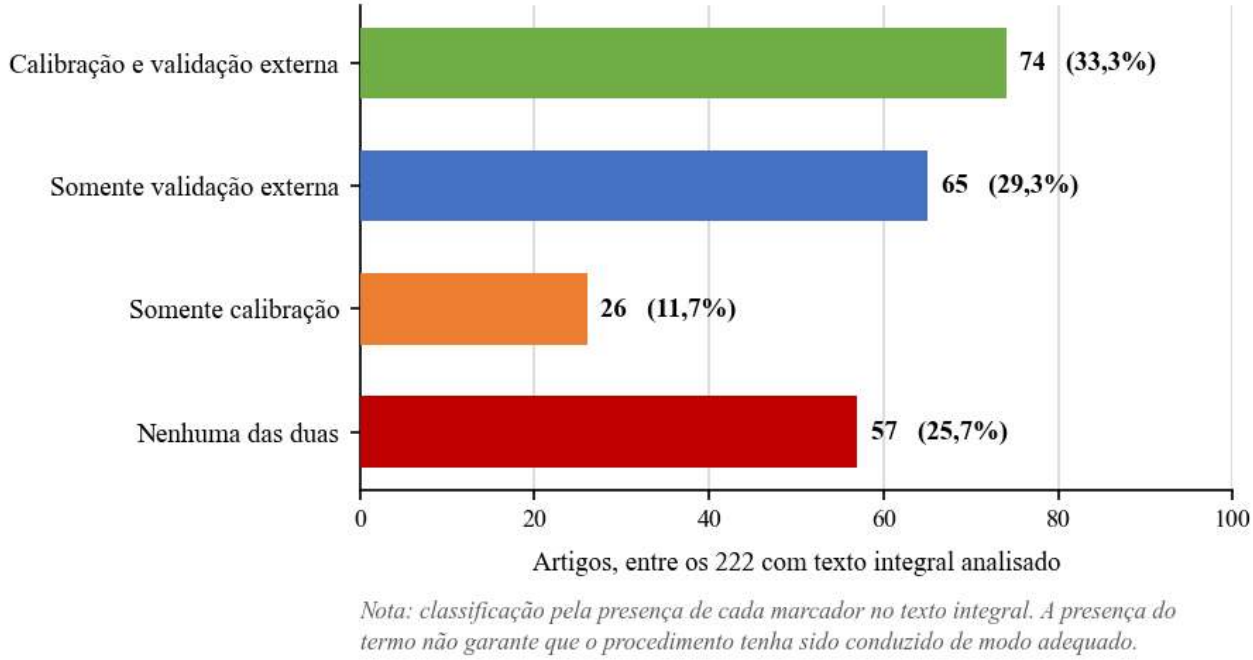
Fonte: elaboração própria a partir do corpus construído no Europe PMC.

O Gráfico 1 traz o achado central, e ele tem duas camadas. A primeira é a distância entre o que o resumo informa e o que o artigo contém. A validação externa é mencionada em 35 resumos, 15,8%,

e em 139 textos integrais, 62,6%: 104 artigos a realizam sem anunciá-la no resumo. A calibração aparece em 28 resumos, 12,6%, e em 100 textos, 45,0%, diferença de 72 artigos. A menção ao TRIPOD passa de 1 resumo, 0,5%, para 41 textos, 18,5%.

A segunda camada está no próprio texto integral. Dos 222 artigos, 100 mencionam calibração, mas apenas 48, ou 21,6%, apresentam curva, gráfico ou inclinação de calibração, que é a forma pela qual a propriedade se demonstra; e 38, ou 17,1%, relatam o escore de Brier. Em outras palavras, metade dos que tocam no assunto o fazem sem apresentar o instrumento correspondente. A distância entre mencionar e demonstrar é, aqui, tão relevante quanto a distância entre o resumo e o texto.

Gráfico 2. Composição dos 222 artigos segundo a presença das duas propriedades no texto integral



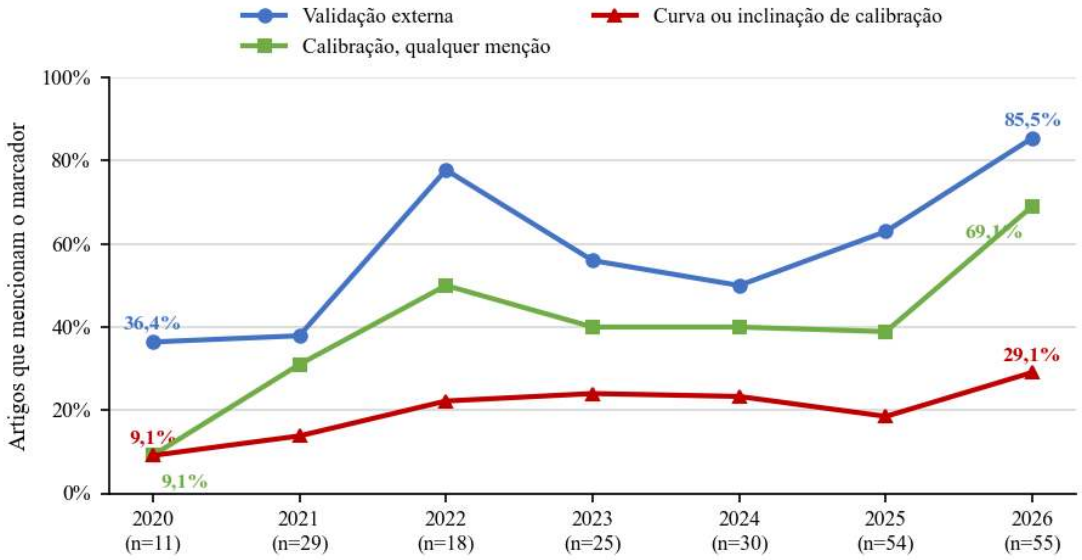
Nota: classificação pela presença de cada marcador no texto integral. A presença do termo não garante que o procedimento tenha sido conduzido de modo adequado.

Fonte: elaboração própria a partir do corpus construído no Europe PMC.

O Gráfico 2 classifica os 222 artigos. São 74, ou 33,3%, os que mencionam calibração e validação externa; 65, ou 29,3%, os que mencionam apenas validação externa; e 26, ou 11,7%, apenas calibração. Outros 57, ou 25,7%, não mencionam nenhuma das duas, embora todos relatem medida de discriminação, por construção do corpus. Um quarto dos estudos brasileiros que avaliam modelos de risco clínico, portanto, apresenta a capacidade de ordenar sem discutir nem a fidelidade das probabilidades nem o comportamento fora da amostra de origem.

3.3 A TENDÊNCIA NO PERÍODO

Gráfico 3. Evolução anual dos marcadores no texto integral, 2020 a 2026



Nota: percentuais calculados sobre os artigos de cada ano entre os 222 com texto integral analisado. O ano de 2026 está incompleto na data da consulta.

Fonte: elaboração própria a partir do corpus construído no Europe PMC.

O Gráfico 3 mostra que o quadro não é estático e que a direção do movimento é favorável. A menção à validação externa passou de 36,4% dos artigos de 2020 para 85,5% dos de 2026, e a menção à calibração, de 9,1% para 69,1%. O ano de 2022 destoa da tendência, com 77,8% de validação externa em apenas 18 artigos, oscilação compatível com o tamanho reduzido do grupo anual.

A curva ou inclinação de calibração, porém, acompanha o movimento em ritmo bem mais lento: de 9,1% em 2020 para 29,1% em 2026. Enquanto a menção à calibração cresceu cerca de sete vezes no período, a apresentação do instrumento que a demonstra cresceu cerca de três. A distância entre as duas linhas, que era de nada em 2020, chega a 40 pontos percentuais em 2026. É nessa distância que reside o problema de relato: o vocabulário da calibração difundiu-se mais depressa que a prática de demonstrá-la.

A menção ao TRIPOD oscila sem tendência clara, entre 9,1% em 2020 e 18,2% em 2026, com pico de 36,0% em 2023. Como esse guia organiza justamente o relato de modelos preditivos, incluindo a exigência de reportar calibração, a sua adoção minoritária é consistente com os demais achados. A governança do uso clínico de algoritmos, incluindo o monitoramento de vieses e erros, tem sido objeto

de atenção crescente na literatura brasileira (Silva *et al.*, 2026; Cardoso *et al.*, 2026), assim como a organização institucional desse acompanhamento nos serviços (Jalacassa; Raimundo; Azevedo, 2026).

4 RESULTADOS E DISCUSSÃO

O primeiro resultado é normativo. A Resolução nº 657/2022, em 19.616 caracteres e 25 artigos, exige validação com evidências suficientes para garantir exatidão, confiabilidade e desempenho, e não emprega uma única vez os termos calibração, discriminação, validação externa, viés, subgrupo ou generalização. O vocabulário é o da engenharia de software, não o da avaliação de modelos preditivos. Para um sistema determinístico, exatidão e repetibilidade descrevem bem o que importa; para um modelo probabilístico treinado em dados, não descrevem.

O segundo resultado é que a literatura brasileira relata muito mais do que os resumos deixam ver. Entre os 222 artigos analisados, a validação externa aparece em 15,8% dos resumos e em 62,6% dos textos, e a calibração em 12,6% e 45,0%. Esse descompasso tem consequência prática: resumos são o que a maior parte dos leitores, e boa parte dos gestores, efetivamente lê, e são o que os mecanismos automatizados de triagem indexam. Um modelo cuja validação externa não apareça no resumo é, para efeitos de busca e de seleção rápida, um modelo sem validação externa.

O terceiro resultado qualifica o segundo e é o mais importante. Mencionar calibração não é demonstrá-la. Dos 100 artigos que tocam no tema no texto integral, apenas 48 apresentam curva, gráfico ou inclinação de calibração e 38 relatam escore de Brier. A detecção empregada aqui é generosa: conta qualquer ocorrência do radical. O número real de estudos com avaliação adequada de calibração é, portanto, no máximo o que se observa, e provavelmente menor.

O quarto resultado é a existência de um quarto do corpus sem nenhuma das duas propriedades. São 57 artigos, 25,7%, que relatam discriminação e não mencionam calibração nem validação externa. São estudos que afirmam que o modelo distingue bem os pacientes, sem afirmar que as probabilidades produzidas são fiéis nem que o desempenho se sustenta fora da amostra de treino. Nada aqui permite concluir que esses modelos sejam inadequados; o que se demonstra é que o relato não permite ao leitor avaliá-los nesses dois aspectos.

O quinto resultado é a melhora consistente. A validação externa passou de 36,4% para 85,5% entre 2020 e 2026, e a calibração de 9,1% para 69,1%. A comunidade científica brasileira incorporou essas exigências com rapidez apreciável, e seria incorreto descrever o quadro como estagnado. A ressalva é que a curva de calibração, instrumento que efetivamente demonstra a propriedade, cresceu de 9,1% para apenas 29,1%, abrindo uma distância de 40 pontos em relação à simples menção.

Esses resultados conversam entre si de modo sugestivo. A norma que autoriza o uso clínico não nomeia as propriedades; a literatura que as discute fá-lo cada vez mais, mas de forma desigual entre mencionar e demonstrar; e o canal de leitura mais usado, o resumo, é justamente onde a informação menos aparece. Não se trata de atribuir à norma a causa do padrão observado na literatura, pois os dois planos têm dinâmicas próprias e este desenho não permite estabelecer relação causal entre eles. Trata-se de registrar que, nos dois planos, o vocabulário disponível para cobrar calibração e transportabilidade é mais escasso do que o uso clínico desses modelos recomendaria. Estudos brasileiros que empregam aprendizado de máquina sobre bases nacionais vêm se multiplicando em domínios diversos (Bezerra; Cunha, 2025; Matos *et al.*, 2026; Rosa, 2025; Santos, 2025), o que torna a questão do relato progressivamente mais relevante.

Do ponto de vista de quem decide incorporar um modelo, a consequência é operacional. Um serviço que avalie adotar uma ferramenta de estratificação de risco encontra, hoje, uma norma que exige validação sem especificar em que consiste, e uma literatura em que a informação necessária existe mais no texto do que no resumo e mais como menção do que como demonstração. A verificação exigida do gestor é, por isso, mais trabalhosa do que precisaria ser, e a questão da transportabilidade entre contextos assistenciais de complexidade distinta já foi apontada como crítica no caso dos modelos de risco (Daniel, 2025).

5 LIMITAÇÕES E LACUNAS

A primeira limitação, e a que mais condiciona a leitura, é que a detecção léxica mede a presença de termos, não a qualidade do procedimento. Um artigo classificado aqui como mencionando calibração pode tê-la avaliado de modo inadequado, e um artigo que a tenha avaliado com rigor pode ter usado terminologia não capturada pelos padrões empregados. Os percentuais devem ser lidos como limites superiores da prática adequada, nunca como medida dela.

A segunda limitação é de corpus. A base utilizada indexa predominantemente periódicos internacionais, de modo que estudos publicados em revistas brasileiras não indexadas no Europe PMC ficaram de fora. O critério de afiliação brasileira inclui artigos de coautoria internacional cuja coorte pode não ser brasileira, e exclui eventuais estudos sobre coortes brasileiras sem autor afiliado no País. O corpus é, portanto, uma aproximação da produção brasileira, não o seu universo.

A terceira limitação é o viés de acesso aberto. Dos 249 artigos com identificador no PubMed Central, obteve-se texto integral de 222, e esse subconjunto é majoritariamente de acesso aberto. Caso periódicos fechados tenham práticas de relato sistematicamente distintas, os percentuais aqui apresentados não se transferem a eles.

A quarta limitação refere-se ao componente normativo. Examinou-se uma resolução, não o conjunto de normas técnicas, guias e instruções normativas que podem preencher os conceitos abertos que ela emprega, nem a prática concreta de avaliação da agência nos processos de regularização. A ausência de um termo no texto da resolução não equivale à ausência da exigência correspondente no procedimento administrativo.

Permanecem como lacunas três questões. Não foi possível determinar, nos artigos que apresentam calibração, se ela foi avaliada na amostra de desenvolvimento ou na de validação externa, distinção metodologicamente decisiva que exigiria leitura individual. Não foi possível verificar quantos dos modelos descritos chegaram a uso clínico efetivo nem quantos foram regularizados como software dispositivo médico. E não foi possível estabelecer se a melhora observada decorre de difusão de padrões de relato, de exigência editorial ou de mudança na composição dos periódicos em que os estudos brasileiros são publicados.

6 CONCLUSÃO

A estratificação de risco com aprendizado de máquina depende de duas propriedades que nenhum índice de discriminação captura: a calibração, que torna a probabilidade utilizável, e a validação externa, que torna o modelo transportável. Este trabalho examinou como as duas aparecem na norma brasileira que regula o software como dispositivo médico e na produção científica com afiliação brasileira entre 2020 e 2026.

No plano normativo, a Resolução da Diretoria Colegiada nº 657/2022 exige validação com exatidão, confiabilidade, desempenho e repetibilidade, e não emprega uma única vez os termos calibração, discriminação, validação externa, viés, subgrupo ou generalização em seus 19.616 caracteres. O vocabulário é o de sistemas de software determinísticos, aplicado a modelos probabilísticos aprendidos de dados.

No plano científico, entre os 222 artigos brasileiros com texto integral analisado que relatam discriminação, 62,6% mencionam validação externa e 45,0% calibração, contra 15,8% e 12,6% nos respectivos resumos. Apenas 21,6% apresentam curva ou inclinação de calibração e 18,5% citam o TRIPOD. Um quarto dos estudos, 25,7%, não menciona nenhuma das duas propriedades.

A tendência, porém, é de melhora nítida: a validação externa passou de 36,4% dos artigos de 2020 para 85,5% dos de 2026, e a calibração de 9,1% para 69,1%. O ponto que não acompanhou esse movimento é a demonstração: a curva de calibração foi de 9,1% para 29,1%, abrindo uma distância de 40 pontos percentuais em relação à simples menção do termo.

ISSN: 2177-3246

A contribuição deste trabalho é oferecer, com método reproduzível e sobre um corpus definido, a medida dessas duas distâncias: a que separa o resumo do texto integral e a que separa mencionar de demonstrar. Nomear calibração e validação externa no texto da norma, exigir que ambas constem do resumo estruturado e adotar o TRIPOD como referência de relato são decisões que não dependem de nova tecnologia nem de nova lei, e que tornariam verificável aquilo que hoje o leitor precisa reconstruir artigo por artigo.

REFERÊNCIAS

APOLINARIO, Priscila Peruzzo; OLIVEIRA, Bruna Andrade de; CARVALHO, Michelle das Neves Lopes; LIMA, Maria Helena de Melo. A inteligência artificial na saúde: o cuidado mediado por algoritmos. Lumen et Virtus, v. 17, n. 64, p. e14700, 2026. DOI: https://doi.org/10.56238/levv17n64-143.

AQUINO, Jaine Fernandes de; FERNANDES, Camily Pires. Análise temporal dos marcadores do consumo alimentar da população adulta em Mato Grosso do Sul: dados do SISVAN, 2018-2022. Aracê, v. 8, n. 9, p. e14501, 2026. DOI: https://doi.org/10.56238/arev8n9-088.

BEZERRA, Jackson Henrique da Silva; CUNHA, Mônica Pereira Lima. Uso de aprendizado de máquina na predição de óbitos em pacientes internados por SRAG: revisão integrativa. Revista Interfaces: Saúde, Humanas e Tecnologia, v. 12, n. 4, 2025. DOI: https://doi.org/10.16891/2317-434x.v12.e4.a2024.pp4688-4705.

BRASIL. Agência Nacional de Vigilância Sanitária. Resolução de Diretoria Colegiada - RDC nº 657, de 24 de março de 2022. Dispõe sobre a regularização de software como dispositivo médico (Software as a Medical Device - SaMD). Diário Oficial da União: seção 1, Brasília, DF, 30 mar. 2022. Disponível em: https://www.in.gov.br. Acesso em: 17 out. 2026.

CARDOSO, Rodrigo Eduardo Rocha; NASCIMENTO, Isabele Pereira; PAULO, Vanessa Souto; LÔPO, Pã da Silva. Inteligência artificial, vulnerabilidade e direito à saúde: riscos da discriminação algorítmica e desafios à proteção jurídica. Lumen et Virtus, v. 17, n. 61, p. e13491, 2026. DOI: https://doi.org/10.56238/levv17n61-044.

DANIEL, Lucas Pedroza. Modelos de risco cirúrgico em contextos de baixa complexidade: validação externa e aplicabilidade na atenção primária. Revista ft, v. 29, n. 153, p. 30-31, 2025. DOI: https://doi.org/10.69849/revistaft/ra10202512171130.

DEGGERONI, Maria. O desafio da regulação da inteligência artificial na saúde: aspectos técnico-jurídicos da segurança das informações. In: Inteligência artificial, proteção de dados e inovação. p. 167-188, 2023. DOI: https://doi.org/10.36592/9786554600552-08.

EUROPE PMC. Base bibliográfica de literatura biomédica: consulta a estudos com afiliação brasileira sobre modelos preditivos com aprendizado de máquina, 2020 a 2026. Hinxton: EMBL-EBI, 2026. Disponível em: https://www.ebi.ac.uk/europepmc. Acesso em: 17 out. 2026.

JALACASSA, Domingos Fernandes; RAIMUNDO, Christopher; AZEVEDO, Caroline Victória Quadros de. A criação obrigatória de comissões de IA e telemedicina nos hospitais: o novo modelo de governança clínica. Lumen et Virtus, v. 17, n. 64, p. e14297, 2026. DOI: https://doi.org/10.56238/levv17n64-008.

JAMBOR, Daniela Guarita. Avaliação regulatória de software como dispositivo médico com foco na inclusão e equidade nas aplicações da inteligência artificial. Revista de Direito Sanitário, v. 25, n. 1, p. e0017, 2025. DOI: https://doi.org/10.11606/issn.2316-9044.rdisan.2025.231708.

MATOS, Paulo José dos Santos de; GOMES FILHO, Isaac Suzart; ARAÚJO, Edna Maria de; CRUZ, Simone Seixas da. Inteligência artificial, envelhecimento e saúde coletiva: aprendizado de máquina no rastreamento de sintomas depressivos. In: Perspectivas multidisciplinares em saúde, p. 838-847, 2026. DOI: https://doi.org/10.47094/978-65-284-0495-7/838-847.

MORAES, Daniele Alves; FERREIRA, Kauany Aparecida Martins. Inteligência artificial e acesso à justiça em demandas de políticas públicas: eficiência judicial, desigualdades e supervisão. Aracê, v. 8, n. 10, p. e14912, 2026. DOI: https://doi.org/10.56238/arev8n10-065.

MORAIS, Rodrigo Martos de. Software como dispositivo médico (SaMD) e a regulação pela Anvisa: desafios e oportunidades. Revista de Geopolítica, v. 16, n. 5, p. e851, 2025. DOI: https://doi.org/10.56238/revgeov16n5-070.

ROSA, Lucio Junio Benfica. Sofrimento emocional no Brasil: uma abordagem de aprendizado de máquina com dados da PNS 2019. In: A saúde pública e o bem-estar da sociedade, p. 36-41, 2025. DOI: https://doi.org/10.47573/aya.5379.2.451.5.

SANTOS, Ana Beatriz Silva. Impacto da inteligência artificial na predição de pré-eclâmpsia: como algoritmos de aprendizado de máquina podem antecipar o diagnóstico. In: Saúde em foco: estudos, pesquisas e aplicações, p. 347-355, 2025. DOI: https://doi.org/10.63330/aurumpub.024-036.

SILVA, Artur Lima da; GOMES, Veronica do Carmo Santos; RAMO, Thaís Fernandes do; SANTOS, Evelyn Fernandes; MAGIOLI, Dilce Maria Mendes. Governança da inteligência artificial na saúde: desafios para o monitoramento de vieses, erros algorítmicos e segurança do paciente. Lumen et Virtus, v. 17, n. 65, p. e14835, 2026. DOI: https://doi.org/10.56238/levv17n65-019.