REVISÃO DE LITERATURA IDENTIFICADO

VIÉS RACIAL E SOCIOECONÔMICO EM MODELOS DE INTELIGÊNCIA ARTIFICIAL APLICADOS À SAÚDE: REVISÃO DE ESCOPO

RACIAL AND SOCIOECONOMIC BIAS IN ARTIFICIAL INTELLIGENCE MODELS APPLIED TO HEALTH: A SCOPING REVIEW

Vitor Comassetto Paes[1]; Gabriel Henrique Carvalho Leite Romeiro[1]; Lavínia Vital Quixadá Souza[1]; Jessyca Lane Fausto Lira[1]; Luciano Timbo Barbosa[1]; Reinaldo dos Santos Moura[2]

1 Acadêmico(a) de Medicina, Centro Universitário de Maceió (UNIMA/Afya), Maceió/AL

2 Docente de Medicina, Centro Universitário de Maceió (UNIMA/Afya), Maceió/AL

E-mail do autor principal: reinaldo.santos@afya.com.br

Eixo Temático: Eixo 03 - Tecnologia e Inovação

RESUMO

Modelos de inteligência artificial (IA) e aprendizado de máquina têm sido incorporados a aplicações clínicas de diagnóstico, predição de risco e alocação de recursos, mas evidências recentes apontam desempenho desigual entre subgrupos populacionais definidos por raça, gênero, renda ou idade, fenômeno conhecido como viés algorítmico. Objetivou-se mapear a extensão, a natureza e as lacunas da literatura sobre viés algorítmico em IA aplicada à saúde. Trata-se de revisão de escopo conduzida segundo a metodologia do Joanna Briggs Institute (JBI) e reportada conforme o checklist PRISMA-ScR, com buscas em PubMed/MEDLINE, LILACS/BVS e SciELO, sem restrição geográfica, e critérios de elegibilidade exigindo métrica quantitativa comparativa de desempenho entre subgrupos, resultando em 321 registros identificados, 292 após remoção de duplicatas, 151 estudos elegíveis e 20 estudos primários incluídos na síntese quantitativa, além de 76 revisões de domínio tratadas como literatura comparadora. Os 20 estudos incluídos foram publicados entre 2024 e 2026, aplicados sobretudo a predição de risco clínico e imagem diagnóstica, com desempenho sistematicamente inferior em subgrupos raciais/étnicos minoritários e em populações de baixa renda na maioria dos estudos. Poucos estudos relataram estratégias de mitigação validadas externamente. Conclui-se que a literatura sobre viés algorítmico em IA na saúde é numerosa e já conta com diversas revisões de enquadramento geral, mas há lacuna de estudos com validação externa multicêntrica e estratégias de mitigação padronizadas, sobretudo em contextos de baixa e média renda.

Palavras-chave: Inteligência Artificial. Viés (Epidemiologia). Equidade em Saúde. Aprendizado de Máquina. Disparidades em Assistência à Saúde.

ABSTRACT

Artificial intelligence (AI) and machine learning models have been incorporated into clinical applications for diagnosis, risk prediction, and resource allocation, but recent evidence points to unequal performance across population subgroups defined by race, gender, income, or age, a phenomenon known as algorithmic bias. The objective was to map the extent, nature, and gaps in the literature on algorithmic bias in health AI. This is a scoping review conducted according to the Joanna Briggs Institute (JBI) methodology and reported according to the

PRISMA-ScR checklist, with searches in PubMed/MEDLINE, LILACS/BVS, and SciELO, without geographic restriction, and eligibility criteria requiring an explicit quantitative comparative performance metric between subgroups, resulting in 321 records identified, 292 after duplicate removal, 151 eligible studies, and 20 primary studies included in the quantitative synthesis, plus 76 domain reviews treated as comparator literature. The 20 included studies were published between 2024 and 2026, applied mainly to clinical risk prediction and diagnostic imaging, with systematically lower performance in racial/ethnic minority subgroups and low-income populations in most studies. Few studies reported externally validated mitigation strategies. It is concluded that the literature on algorithmic bias in health AI is extensive and already includes several general framing reviews, but there is a gap in multicenter externally validated studies and standardized mitigation strategies, especially in low- and middle-income contexts.

Keywords: Artificial Intelligence. Bias. Health Equity. Machine Learning. Healthcare Disparities.

INTRODUÇÃO

Modelos de IA e aprendizado de máquina vêm sendo incorporados aos sistemas de saúde para apoiar decisões clínicas, diagnósticos, estratificação de risco e gestão de recursos. Entretanto, por serem treinados com dados historicamente produzidos em contextos marcados por desigualdades sociais e assistenciais, esses sistemas podem reproduzir e até ampliar disparidades já existentes, resultando em desempenho desigual entre grupos populacionais definidos por características como raça/cor, gênero, idade e condição socioeconômica. Esse fenômeno é amplamente reconhecido na literatura como viés algorítmico[1].

A preocupação com o viés em IA transcende discussões teóricas: estudos demonstram menor acurácia diagnóstica em populações sub-representadas, subestimação de riscos e dificuldade de generalização para grupos historicamente vulnerabilizados, o que compromete a qualidade do cuidado e aprofunda iniquidades já presentes nos sistemas de saúde[1].

Nos últimos anos, diversas revisões abordaram os impactos da IA sobre as desigualdades em saúde, incluindo análises na Atenção Primária[2], disparidades raciais associadas ao uso de sistemas inteligentes[3], estratégias de mitigação de vieses em dados do mundo real[4] e lacunas na representação sociodemográfica em estudos de IA aplicada à oftalmologia[5], além de protocolo recente para mapear viés na interpretação automatizada de eletrocardiogramas[6]. Apesar desses avanços, ainda há escassez de sínteses focadas em estudos primários que avaliem comparativamente o desempenho entre subgrupos. Diante dessa lacuna, esta revisão teve como objetivo mapear a extensão, as características e as lacunas da literatura sobre viés algorítmico e desempenho desigual em sistemas de IA na saúde.

MÉTODO

A revisão de escopo foi conduzida de acordo com a metodologia da JBI e reportada conforme o PRISMA-ScR. A questão foi estruturada em formato PCC: população de pacientes ou usuários de saúde submetidos a IA/aprendizado de máquina; conceito de viés algorítmico e desempenho desigual entre subgrupos, mensurado por métrica quantitativa comparativa; contexto de aplicações clínicas em saúde, sem restrição de idioma ou geografia. O escopo inicial (desigualdades regionais em saúde no Brasil) foi ampliado para o contexto global por escassez de estudos, resultando em mais de 1.500 registros elegíveis; optou-se por refinar o conceito para viés algorítmico, refinamento iterativo compatível com a orientação da JBI.

As buscas foram realizadas em PubMed/MEDLINE, LILACS/BVS e SciELO (Tabela 1), com estratégia ampla seguida de filtragem por descritores de viés/equidade/desempenho por subgrupo; o baixo retorno da SciELO (n=4) é limitação declarada. Incluíram-se estudos primários e revisões sobre IA em saúde com evidência de desempenho diferencial entre grupos; excluíram-se publicações sem relação com IA ou viés/equidade, editoriais, estudos fora da saúde e trabalhos conceituais. Para os primários, exigiu-se ainda métrica quantitativa por subgrupo, publicação a partir de 2024 e amostra relatada.

A seleção ocorreu por título e resumo, por revisor único; a leitura integral ficou restrita a casos de ambiguidade em acesso aberto, e as revisões foram segregadas como literatura comparadora. A busca identificou 321 registros (280 PubMed/MEDLINE, 37 LILACS/BVS, 4 SciELO); após remoção de 29 duplicatas, 292 avançaram à triagem, que excluiu 141, restando 151 elegíveis (76 revisões e 75 primários). Os critérios adicionais (métrica quantitativa, ano, amostra) reduziram o conjunto a 20 estudos incluídos na síntese final (Figura 1); as 76 revisões, mais 5 relevantes e 1 estudo fundacional sobre justiça distributiva, compuseram a literatura comparadora.

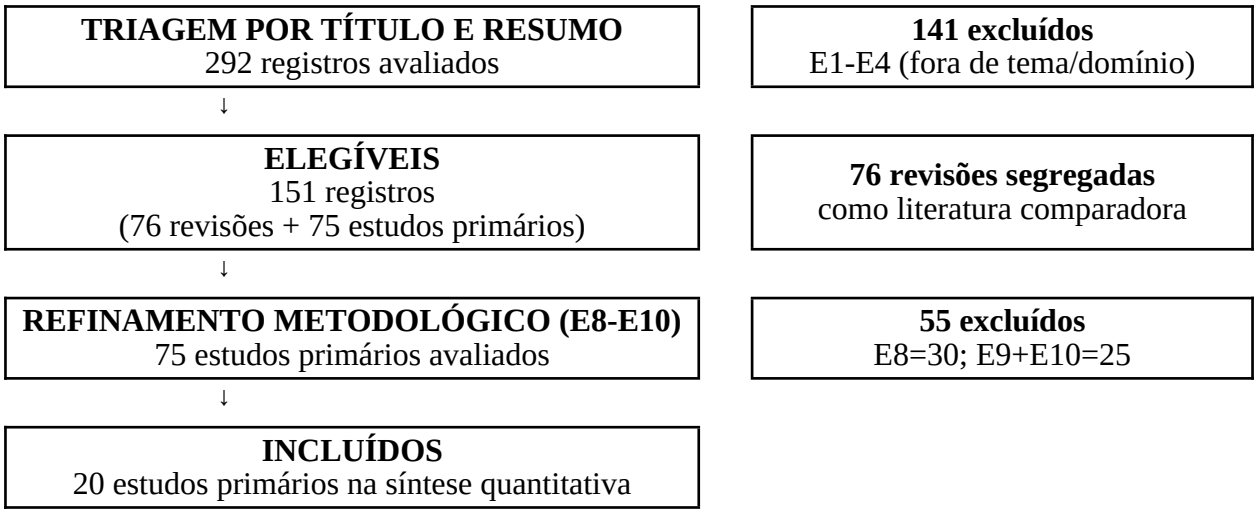
Base	Estratégia (retrieval amplo + filtro programático por viés/equidade)
PubMed/MEDLINE	("artificial intelligence"[tiab] OR "machine learning"[tiab] OR "deep learning"[tiab] OR algorithm*[tiab]) AND (health[tiab] OR clinical[tiab] OR patient*[tiab] OR medical[tiab]); filtro pós-hoc: bias OR fairness OR disparit* OR equity OR subgroup
LILACS/BVS	"inteligência artificial" AND saúde, sem filtro geográfico; filtro pós-hoc: viés OR equidade OR disparidade OR desempenho por subgrupo
SciELO	"inteligência artificial" AND saúde; filtro pós-hoc idêntico ao LILACS

Tabela 1. Estratégias de busca por base de dados, sem restrição de data ou idioma. Filtro programático aplicado sobre título e resumo dos registros recuperados.

Figura 1. Fluxograma PRISMA-ScR.

IDENTIFICAÇÃO
321 registros identificados (PubMed 280; LILACS 37; SciELO 4)
29 duplicatas removidas, restando 292 registros

↓



Nota: E1 ausência de termo de IA/aprendizado de máquina; E2 ausência de sinal de viés/equidade/subgrupo; E3 editorial/carta/comentário/notícia; E4 domínio fora da saúde; E8 ausência de métrica quantitativa explícita por subgrupo; E9 publicação anterior a 2024; E10 tamanho amostral não relatado.

RESULTADOS

Os 20 estudos incluídos (Figura 1, Tabela 2) foram publicados entre 2024 e 2026, majoritariamente em 2025-2026, com origem predominante nos Estados Unidos, além de Canadá, Israel, Gana e México. Predominaram desenhos retrospectivos de coorte única com validação interna; poucos tiveram validação externa multicêntrica.

Em predição de risco clínico, sete estudos avaliaram desempenho diferencial entre subgrupos: readmissão em ICC com equidade melhorada por determinantes sociais[7], fibrilação atrial com escore livre de raça superior ao CHARGE-AF[8], transtornos de humor perinatal com mitigação por reponderação[9], doença renal crônica com modelo modular por raça superior ao cego à raça[10], reabilitação pós-cirúrgica em escoliose pediátrica com calibração para IA responsável[11], depressão pré-natal em população negra/latina com desempenho modesto[12] e declínio cognitivo no México com viés em subgrupos sub-representados[13].

Em imagem e dados clínicos, quatro estudos: retinopatia diabética em indígenas australianos com sensibilidade superior à do especialista[14], estadiamento de câncer de mama com desempenho superior em brancas mesmo após calibração[15], seleção para angiografia coronariana com validação externa geo/temporal[16] e detecção de glaucoma por notas clínicas com disparidade racial[17].

Em registros digitais, três estudos identificaram viés na coleta: frequência de mensuração desigual em prontuário de UTI por raça/gênero[18], viés por tipo de dispositivo em

avaliação remota de Parkinson (possível proxy socioeconômico)[19] e mais falsos positivos de não adesão em pacientes negros e de baixa renda[20].

Em alocação de recursos, três estudos trataram absenteísmo em atenção primária sem viés por sexo/raça[21], roteamento digital de urgência com implicações de equidade[22] e teste de LDL-C sem jejum para telessaúde[23]. Outros três: mitigação de injustiça em sobrevida de câncer de bexiga[24], disparidade em metade dos modelos de COVID-19 pediátrico, parcialmente mitigada[25], e arcabouço de IA justa para pré-natal em Gana[26].

Em síntese, desempenho inferior em subgrupos raciais/étnicos minoritários e de baixa renda foi relatado na maioria dos estudos; mitigação foi testada em nove, com resultados mistos, e validação externa multicêntrica foi rara.

Autor/Ano	Aplicação clínica	N	Achado principal (subgrupo)
Fensore C et al., 2026	Readmissão em ICC	33.579	SDOH ampliados melhoram AUC (0,671) e eOR racial (0,437).
Segar MW et al., 2026	Predição de fibrilação atrial	~30.647	Escore livre de raça supera CHARGE-AF (C-index 0,83).
Wong EF et al., 2024	Transtornos de humor perinatal	NI	Reponderação reduz disparidade demográfica entre grupos raciais.
Behmard N et al., 2025	Progressão de doença renal crônica	~30.000	Modelo modular por raça supera modelo cego à raça em acurácia.
Shi W et al., 2025	Reabilitação pós-cirurgia (escoliose)	455	Calibração de IA responsável entre subgrupos pré-operatórios.
Huang Y et al., 2024	Depressão pré-natal (EMR)	5.875	AUROC modesto (0,61) em mulheres negras/latinas de baixa renda.
He Y et al., 2026	Declínio cognitivo (México)	4.095	Viés em subgrupos sub-representados (RMSE 39,25).
Chia MA et al., 2024	Retinopatia diabética (indígenas)	864	Sensibilidade do sistema (98,0%) superior ao especialista (87,1%).
Soltan A, Washington P, 2024	Estadiamento de câncer de mama	842	Desempenho superior em pacientes brancas mesmo após calibração.
Petch J et al., 2026	Seleção para angiografia coronariana	319.012	Validação geográfica e temporal externas; equidade por subgrupo.
Mohammadjafari A et al., 2025	Detecção de glaucoma (notas clínicas)	10.000	AUC 0,80; disparidade racial de desempenho entre grupos.
Shi J et al., 2025	Viés de mensuração em UTI (EHR)	>40.000	Frequência de coleta de dados desigual por raça/gênero.
Tumpa ZN et al., 2025	Avaliação remota de Parkinson	251	Viés por tipo de dispositivo e lateralidade manual, não por sexo/etnia.
Mukhopadhyay A et al., 2026	Adesão medicamentosa (ICC)	34.697	Mais falsos positivos em pacientes negros e de baixa renda.
Tuan W et al., 2025	Absenteísmo em atenção primária	109.328 pac.	Sem viés relevante por sexo/raça (AUC 0,85-0,92).
Bashkin O et al., 2026	Roteamento digital de urgência	77.508	Disparidades em taxa de aprovação de solicitações de atendimento.
Doku R et al., 2025	Teste de LDL-C sem jejum	3.477	Reduz barreira estrutural de acesso; equidade racial avaliada.
Carbunaru S et al., 2024	Sobrevida em câncer de bexiga	16.481	Mitigação de injustiça melhora razão de chances equalizadas.
Libin A et al., 2024	Desfecho de COVID-19 pediátrico	9.935	Disparidade subgrupal em 50,7% dos modelos; dados sintéticos mitigam parte.
Osborne A, Usani K, 2025	Cobertura de pré-natal (Gana)	3.314	Arcabouço de IA justa com explicabilidade e mitigação de viés em país de baixa renda.

Tabela 2. Síntese dos 20 estudos primários incluídos, em ordem de citação no texto. NI = não informado; ICC = insuficiência cardíaca congestiva; eOR = razão de chances equalizadas.

CONCLUSÃO

A literatura recente demonstra que o desempenho desigual de sistemas de IA em saúde é fenômeno frequente, sobretudo em predição de risco, imagem médica e saúde perinatal. Apesar do crescimento das evidências, permanecem escassas as estratégias de mitigação com validação externa e os estudos de baixa e média renda, reforçando a necessidade de modelos mais transparentes e equitativos. Os achados devem ser interpretados à luz da triagem por revisor único e da restrição das buscas a bases de dados disponíveis.

REFERÊNCIAS

1. Rajkomar A, Hardt M, Howell MD, Corrado G, Chin MH. Ensuring Fairness in Machine Learning to Advance Health Equity. Ann Intern Med. 2018;169(12):866-872.

2. d'Elia A, Gabbay M, Rodgers S, Kierans C, Jones E, Durrani I, et al. Artificial intelligence and health inequities in primary care: a systematic scoping review and framework. Fam Med Community Health. 2022;10(Suppl 1):e001670.
3. Haider SA, Borna S, Gomez-Cabello CA, Pressman SM, Haider CR, Forte AJ. The Algorithmic Divide: A Systematic Review on AI-Driven Racial Disparities in Healthcare. J Racial Ethn Health Disparities. 2026;13(1):188-217.
4. Huang Y, Guo J, Chen WH, Lin HY, Tang H, Wang F, et al. A scoping review of fair machine learning techniques when using real-world data. J Biomed Inform. 2024;151:104622.
5. Bondok M, Selvakumar R, Asdo A, Naderi B, Zhang C, Wong C, et al. Sociodemographic Reporting in Artificial Intelligence Studies of Retinal Diseases: A Critical Appraisal of the Literature. Ophthalmol Retina. 2025;9(12):1133-1139.
6. Lawal L, Paton C, English M, Holthof B, Preston T. Mapping Algorithmic Bias in AI-Powered Electrocardiogram Interpretation Across the AI Life Cycle: Protocol for a Scoping Review. JMIR Res Protoc. 2026;15:e82486.
7. Fensore C, Deshpande A, Carrillo-Larco RM, Patel SA, Ho JC. Beyond Composite Indices: Comprehensive Social Determinants Improve Heart Failure Readmission Prediction. J Am Heart Assoc. 2026;15(7):e043735.
8. Segar MW, Keshvani N, Jaeger B, Rosenblatt A, Razavi M, Saeed M, et al. Development and validation of a novel machine learning-based algorithm to predict incident atrial fibrillation: A multicohort analysis. Heart Rhythm. 2026;23(4):e537-e546.
9. Wong EF, Saini AK, Accortt EE, Wong MS, Moore JH, Bright TJ. Evaluating Bias-Mitigated Predictive Models of Perinatal Mood and Anxiety Disorders. JAMA Netw Open. 2024;7(12):e2438152.
10. Behmard N, Koshechkin K, Al-Alwani YM, Schier A, Mahdi M. Artificial Intelligence-Driven Risk Stratification in Chronic Kidney Disease Progression: Minimizing Bias via Race-Specific Algorithms. Cureus. 2025;17(12):e98319.
11. Shi W, Giuste FO, Zhu Y, Tamo BJ, Nnamdi MC, Hornback A, et al. Predicting pediatric patient rehabilitation outcomes after spinal deformity surgery with artificial intelligence. Commun Med (Lond). 2025;5(1):1.
12. Huang Y, Alvernaz S, Kim SJ, Maki P, Dai Y, Peñalver Bernabé B. Predicting Prenatal Depression and Assessing Model Bias Using Machine Learning Models. Biol Psychiatry Glob Open Sci. 2024;4(6):100376.
13. He Y, Leng Y, Vranceanu A, Ritchie CS, Blacker D, Das S. A predictive model for cognitive decline using social determinants of health. JAR Life. 2026;15:100056.
14. Chia MA, Hersch F, Sayres R, Bavishi P, Tiwari R, Keane PA, et al. Validation of a deep learning system for the detection of diabetic retinopathy in Indigenous Australians. Br J Ophthalmol. 2024;108(2):268-273.
15. Soltan A, Washington P. Challenges in Reducing Bias Using Post-Processing Fairness for Breast Cancer Stage Classification with Deep Learning. Algorithms. 2024;17(4).
16. Petch J, Di S, Watson T, Becerra MP, Rosella LC, Natarajan MK, et al. An artificial intelligence prediction model for optimizing patient selection for cardiac imaging for the investigation of suspected coronary artery disease. Eur Heart J Digit Health. 2026;7(3):ztag040.
17. Mohammadjafari A, Lin M, Shi M. Deep Learning-Based Glaucoma Detection Using Clinical Notes: A Comparative Study of Long Short-Term Memory and Convolutional Neural Network Models. Diagnostics (Basel). 2025;15(7).
18. Shi J, Hubbard AE, Fong N, Pirracchio R. Implicit bias in ICU electronic health record data: measurement frequencies and missing data rates of clinical variables. BMC Med Inform Decis Mak. 2025;25(1):241.
19. Tumpa ZN, Zawad MRS, Sollis L, Parab S, Chen IY, Washington P. Quantifying device type and handedness biases in a remote Parkinson's disease AI-powered assessment. NPJ Digit Med. 2025;8(1):550.
20. Mukhopadhyay A, Zhao Y, Chunara R, Kronish IM, Lawrence S, Blecker S, et al. The importance of clinical context in evaluating algorithmic fairness: insights from a medication adherence prediction algorithm. J Am Med Inform Assoc. 2026.
21. Tuan W, Yan Y, Abou Al Ardat B, Felix T, Chen Q. Predicting Missed Appointments in Primary Care: A Personalized Machine Learning Approach. Ann Fam Med. 2025;23(4):294-301.
22. Bashkin O, Shalom T, Yehoshua I, Adler L. Digital emergency routing: analysis of feasibility, utilization, and equity implications. Isr J Health Policy Res. 2026;15(1).
23. Doku R, Osafo NY, Kwagyan J, Southerland WM. Nonfasting, Telehealth-Ready LDL-C Testing With Machine Learning to Improve Cardiovascular Access and Equity. medRxiv. 2025.
24. Carbunaru S, Neshatvar Y, Do H, Murray K, Ranganath R, Nayan M. Survival After Radical Cystectomy for Bladder Cancer: Development of a Fair Machine Learning Model. JMIR Med Inform. 2024;12:e63289.
25. Libin A, Treitler JT, Vasaitis T, Shao Y. Evaluating and Reducing Subgroup Disparity in AI Models: An Analysis of Pediatric COVID-19 Test Outcomes. medRxiv. 2024.
26. Osborne A, Usani K. A fairness-aware machine learning framework for maternal health in Ghana: integrating explainability, bias mitigation, and causal inference for ethical AI deployment. BioData Min. 2025;19(1):3.