

GESTÃO E AUDITORIA DE IA NA SAÚDE: MODELOS DE GOVERNANÇA PARA MONITORAR VIESES E ERROS NA TRIAGEM DE PACIENTES

AI MANAGEMENT AND AUDITING IN HEALTHCARE: GOVERNANCE MODELS TO MONITOR BIASES AND ERRORS IN PATIENT TRIAGE.

GESTIÓN Y AUDITORÍA DE LA IA EN LA ATENCIÓN SANITARIA: MODELOS DE GOBERNANZA PARA EL SEGUIMIENTO DE SESGOS Y ERRORES EN LA CLASIFICACIÓN DE PACIENTES

Maria Tarcila Rabelo Pinheiro¹, Jordana Maria Neiva Barroso Queiroz², Thiago Coelho Ribeiro³, João Felipe Moraes do Nascimento⁴, Maria Valdirene Fernandes Oliveira⁵, Vera Lúcia Bessa de Oliveira⁶, Antonia Diva Ferreira de Sousa⁷, Ana Lúcia Almeida de Sousa⁸, Adriano José Almeida Maia⁹, Ana Karine Silva da Rocha Arrais¹⁰, Emanuele Nascimento Silva¹¹, Francisco José Bezerra da Silva¹²

DOI: 10.54899/dcs.v23i96.7009

Recibido: 11/08/2026 | **Aceptado:** 04/09/2026 | **Publicación en línea:** 11/09/2026.

RESUMO

Objetivo: Analisar modelos de governança e auditoria aplicáveis à IA na saúde, com ênfase na identificação de vieses, erros preditivos, transparência, validação e supervisão humana, buscando ainda descrever mecanismos capazes de fortalecer a segurança e a equidade das decisões

¹ Mestranda em Gestão da Saúde, Universidade Estadual do Ceará (UECE), Fortaleza, Ceará, Brasil.

E-mail: tarcilaenf@yahoo.com.br Orcid: <https://orcid.org/0009-0003-8309-9615>

² Especialista em Enfermagem em Terapia Intensiva, Faculdade Farias Brito, Fortaleza, Ceará, Brasil.

E-mail: enfajordanaqueiroz@gmail.com Orcid: <https://orcid.org/0009-0004-2730-5266>

³ Especialista em UTI Adulto, Centro Universitário Maurício de Nassau (UNINASSAU), Fortaleza, Ceará, Brasil.

E-mail: thiago.cribeiro1911@gmail.com Orcid: <https://orcid.org/0009-0003-5580-6503>

⁴ Especialista em Urgência e Emergência, Faculdade Terra Nordeste (FATENE), Caucaia, Ceará, Brasil.

E-mail: felpemoraes@gmail.com Orcid: <https://orcid.org/0009-0004-0236-9992>

⁵ Especialista em Enfermagem em Terapia Intensiva, Faculdade de Quixeramobim (UNIQ), Quixeramobim, Ceará, Brasil. E-mail: mariavaldirene584@gmail.com Orcid: <https://orcid.org/0009-0006-6260-8515>

⁶ Especialista em Saúde Pública da Família e do Idoso, Faculdade Paulista de Serviço Social (FAPSS), São Caetano do Sul, São Paulo, Brasil. E-mail: vera.bessa9273@gmail.com Orcid: <https://orcid.org/0009-0009-4408-8212>

⁷ Especialista em Urgência e Emergência, Faculdade Ateneu, Fortaleza, Ceará, Brasil.

E-mail: divaferreira2306@gmail.com Orcid: <https://orcid.org/0009-0002-6759-6702>

⁸ Especialista em Enfermagem Neonatal e Pediátrica, Faculdade de Quixeramobim (UNIQ), Fortaleza, Ceará, Brasil. E-mail: anadsousachagas@hotmail.com Orcid: <https://orcid.org/0009-0004-6608-1365>

⁹ Especialista em Saúde Pública e Vigilância Sanitária, Faculdade Venda Nova do Imigrante (FAVENI), Venda Nova do Imigrante, Espírito Santo, Brasil. E-mail: adrianojalmeida2026@gmail.com

Orcid: <https://orcid.org/0009-0006-3260-3238>

¹⁰ Especialista em Linhas de Cuidado em Enfermagem, Universidade Federal de Santa Catarina, Fortaleza, Ceará, Brasil. E-mail: karinerochaarrais@gmail.com Orcid: <https://orcid.org/0009-0004-4592-1740>

¹¹ Especialista em UTI Pediátrica e Neonatal, Centro Universitário Cidade Verde (UNICIVE), Maringá, Paraná, Brasil. E-mail: manuewal@gmail.com Orcid: <https://orcid.org/0009-0000-0054-7063>

¹² Especialista em Enfermagem do Trabalho, Universidade Estadual do Ceará (UECE), Fortaleza, Ceará, Brasil. E-mail: bezerraaph01@gmail.com Orcid: <https://orcid.org/0009-0008-0003-7424>

assistenciais. Métodos: Revisão narrativa da literatura, de abordagem descritiva e analítica, realizada na Biblioteca Virtual em Saúde, incluindo LILACS e MEDLINE, SciELO e PubMed. Utilizaram-se os descritores DeCS/MeSH Inteligência Artificial (*Artificial Intelligence*), Triage (*Triage*), Segurança do Paciente (*Patient Safety*), Equidade em Saúde (*Health Equity*) e Sistemas de Apoio à Decisão Clínica (*Clinical Decision Support Systems*), combinados pelos operadores AND e OR. Após leitura de títulos, resumos e textos completos, sete publicações compuseram o corpus final, analisado de forma descritiva, interpretativa e temática. Resultados: A acurácia dos sistemas de IA na triagem variou entre 80,5% e 99,1%, enquanto o GPT-4 apresentou 66,7% de acurácia na classificação de triagem. A qualidade e a completude dos dados influenciaram o desempenho preditivo, e a avaliação institucional mostrou-se desigual, pois 65% dos hospitais utilizavam IA ou modelos preditivos, mas apenas 61% avaliavam localmente a acurácia e 44% o viés. Também foram identificadas recomendações voltadas à equidade, transparência, explicabilidade, rastreabilidade, monitoramento contínuo e responsabilização. Conclusão: A utilização segura da IA na triagem exige governança permanente, validação local, auditorias periódicas, avaliação do desempenho entre grupos populacionais, qualidade dos dados e supervisão humana, articulando desempenho técnico, responsabilidade institucional e proteção do paciente.

Palavras-chave: Equidade em Saúde. Inteligência Artificial. Segurança do Paciente. Sistemas de Apoio à Decisão Clínica. Triage.

ABSTRACT

Objective: To analyze governance and audit models applicable to AI in healthcare, emphasizing the identification of biases, predictive errors, transparency, validation, and human oversight, while also seeking to describe mechanisms capable of strengthening the safety and equity of care-related decisions. Methods: A narrative literature review with a descriptive and analytical approach was conducted using the Virtual Health Library (including LILACS and MEDLINE), SciELO, and PubMed. The DeCS/MeSH descriptors used were Artificial Intelligence, Triage, Patient Safety, Health Equity, and Clinical Decision Support Systems, combined using the operators AND and OR. After reviewing titles, abstracts, and full texts, seven publications formed the final corpus, which was analyzed using descriptive, interpretive, and thematic methods. Results: The accuracy of AI systems in triage ranged from 80.5% to 99.1%, while GPT-4 showed 66.7% accuracy in triage classification. Data quality and completeness influenced predictive performance, and institutional assessment was uneven: 65% of hospitals used AI or predictive models, yet only 61% locally assessed accuracy and 44% assessed bias. Recommendations regarding equity, transparency, explainability, traceability, continuous monitoring, and accountability were also identified. Conclusion: The safe use of AI in triage requires ongoing governance, local validation, periodic audits, performance assessment across population groups, data quality, and human oversight, integrating technical performance, institutional responsibility, and patient protection.

Keywords: Health Equity. Artificial Intelligence. Patient Safety. Clinical Decision Support Systems. Triage.

RESUMEN

Objetivo: Analizar los modelos de gobernanza y auditoría aplicables a la IA en la atención médica, con énfasis en la identificación de sesgos, errores predictivos, transparencia, validación y supervisión humana, al tiempo que se describen mecanismos capaces de fortalecer la seguridad y la equidad de las decisiones en la atención médica. **Métodos:** Se realizó una revisión narrativa de la literatura, con un enfoque descriptivo y analítico, en la Biblioteca Virtual en Salud, incluyendo LILACS y MEDLINE, SciELO y PubMed. Se utilizaron los descriptores DeCS/MeSH Inteligencia Artificial, Triage, Seguridad del Paciente, Equidad en Salud y Sistemas de Apoyo a la Decisión Clínica, combinados con los operadores AND y OR. Después de leer los títulos, resúmenes y textos completos, siete publicaciones conformaron el corpus final, analizado de forma descriptiva, interpretativa y temática. **Resultados:** La precisión de los sistemas de IA en el triaje osciló entre el 80,5% y el 99,1%, mientras que GPT-4 mostró una precisión del 66,7% en la clasificación del triaje. La calidad y la exhaustividad de los datos influyeron en el rendimiento predictivo, y la evaluación institucional resultó desigual, ya que el 65 % de los hospitales utilizaban IA o modelos predictivos, pero solo el 61 % evaluaba localmente la precisión y el 44 % evaluaba el sesgo. También se identificaron recomendaciones centradas en la equidad, la transparencia, la explicabilidad, la trazabilidad, la monitorización continua y la rendición de cuentas. **Conclusión:** El uso seguro de la IA en el triaje requiere una gobernanza continua, validación local, auditorías periódicas, evaluación del rendimiento en diferentes grupos de población, calidad de los datos y supervisión humana, articulando el rendimiento técnico, la responsabilidad institucional y la protección del paciente.

Palabras clave: Equidad en la Salud. Inteligencia Artificial. Seguridad del Paciente. Sistemas de Apoyo a la Toma de Decisiones Clínicas. Triage.



Esta obra está bajo una [Licencia Creative Commons Atribución- NoComercial 4.0 Internacional](https://creativecommons.org/licenses/by-nc/4.0/)

INTRODUÇÃO

A inteligência artificial (IA) vem ampliando sua presença nos serviços de saúde por meio de sistemas capazes de processar dados clínicos, reconhecer padrões e oferecer suporte à tomada de decisão, alcançando áreas assistenciais marcadas por elevada demanda e necessidade de respostas rápidas. Sua incorporação exige, contudo, atenção às implicações técnicas, éticas e regulatórias relacionadas à segurança do paciente e à utilização de informações sensíveis (Cappato *et al.*, 2025).

No âmbito da gestão em saúde, a adoção da IA ultrapassa a automação de tarefas e passa a interferir na organização dos processos assistenciais, no planejamento institucional e nos mecanismos utilizados para apoiar decisões clínicas. Essa expansão tecnológica requer

infraestrutura adequada, capacitação profissional, protocolos definidos e formas de governança capazes de estabelecer responsabilidades e limites para o emprego dos algoritmos nos serviços de saúde (Hirschmann *et al.*, 2025).

A triagem de pacientes constitui uma das aplicações de maior relevância nesse contexto, pois envolve a classificação da gravidade, o estabelecimento de prioridades e o direcionamento do atendimento em ambientes sujeitos à sobrecarga. Sistemas de IA podem incorporar sinais vitais, sintomas, histórico clínico e outras informações disponíveis para auxiliar a avaliação inicial, sem afastar a necessidade de julgamento clínico e acompanhamento profissional (Cappato *et al.*, 2025).

A dimensão quantitativa dessas aplicações demonstra a relevância de acompanhar não apenas a disponibilidade da tecnologia, mas também seu desempenho clínico. Em sistema de triagem pediátrica baseado em IA, foi registrada acurácia de 94,8%, sensibilidade de 92,5% para identificação de casos de alto risco e redução do tempo médio de espera de 42 para 18 minutos, parâmetros que ilustram a capacidade de mensuração objetiva dessas ferramentas (Hariharan *et al.*, 2026).

Entretanto, índices globais de acurácia podem ocultar falhas clinicamente relevantes quando determinados grupos ou condições aparecem em menor frequência nas bases utilizadas para treinamento. Em uma aplicação preditiva analisada em saúde pública, o modelo inicial deixou de identificar 66 de 72 casos malignos, apresentando sensibilidade de apenas 8,33%, situação que exemplifica a importância de avaliar falsos negativos e desempenho por classes, e não somente a acurácia geral (Silva *et al.*, 2026).

O viés algorítmico pode decorrer da composição das bases de dados, da seleção das variáveis, dos critérios utilizados para definir os desfechos e das diferenças entre a população empregada no desenvolvimento do modelo e aquela atendida no serviço. Na triagem, essas limitações podem produzir classificações inadequadas, restringir a capacidade de generalização e favorecer decisões desiguais quando grupos populacionais permanecem sub-representados nos dados utilizados pelo sistema (Daniel, 2023).

Essa preocupação adquire maior relevância em sistemas públicos de saúde caracterizados por desigualdades territoriais, socioeconômicas, raciais e digitais, nas quais a disponibilidade e a qualidade das informações não se distribuem de maneira homogênea. Dados incompletos, baixa explicabilidade e ausência de avaliação por subgrupos podem comprometer a equidade, tornando necessário verificar o desempenho da IA conforme características populacionais e contextos

locais de utilização (Lovadini *et al.*, 2026).

A governança da IA em saúde deve, portanto, estabelecer princípios que orientem todo o ciclo de utilização da tecnologia, desde a seleção e validação do modelo até sua aplicação e acompanhamento contínuo. A Organização Mundial da Saúde (OMS), juntamente com referenciais regulatórios internacionais, enfatiza aspectos como transparência, equidade, responsabilização, evidência clínica, supervisão humana e monitoramento de tecnologias digitais utilizadas em situações capazes de influenciar decisões assistenciais (Daniel, 2023).

Nesse cenário, a auditoria algorítmica representa um componente da governança destinado a verificar se o desempenho observado inicialmente permanece adequado após a implantação da ferramenta. Validação clínica local, controle de vieses, governança de dados, integração aos sistemas de informação, capacitação das equipes e monitoramento pós-implantação constituem requisitos que possibilitam identificar alterações de desempenho, erros recorrentes e riscos relacionados ao uso da IA na rotina assistencial (Daniel, 2023).

A gestão dessas tecnologias também pressupõe delimitação de responsabilidades entre instituições, profissionais e desenvolvedores, especialmente quando recomendações produzidas por algoritmos interferem na definição da prioridade clínica. Planejamento estratégico, protocolos institucionais claros e governança ética tornam-se necessários para que a IA permaneça integrada à prática humana como recurso de apoio, evitando que decisões automatizadas substituam indiscriminadamente a avaliação profissional ou fragilizem mecanismos de segurança assistencial (Hirschmann *et al.*, 2025).

A realização deste estudo justifica-se pela necessidade de aproximar a expansão da IA na triagem de mecanismos permanentes de controle capazes de reconhecer erros que podem permanecer invisíveis em indicadores agregados, sobretudo quando atingem populações específicas. A criação de estruturas de governança orientadas por validação local, acompanhamento da equidade, supervisão humana e responsabilização institucional torna-se relevante para conciliar inovação tecnológica, segurança do paciente e qualidade do cuidado (Lovadini *et al.*, 2026).

O problema de pesquisa que orienta este estudo concentra-se na necessidade de compreender como modelos de governança e auditoria podem favorecer o monitoramento de vieses e erros produzidos por sistemas de IA empregados na triagem de pacientes. Assim, objetiva-se analisar modelos de governança e auditoria aplicáveis à IA na saúde, com ênfase na identificação de vieses, erros preditivos, transparência, validação e supervisão humana, buscando

ainda descrever mecanismos capazes de fortalecer a segurança e a equidade das decisões assistenciais.

METODOLOGIA

Trata-se de uma revisão narrativa da literatura, de abordagem descritiva e analítica, desenvolvida com a finalidade de reunir e interpretar produções científicas relacionadas à gestão, governança e auditoria de sistemas de IA aplicados à triagem de pacientes. A escolha desse delineamento possibilitou integrar pesquisas com diferentes desenhos metodológicos, incluindo estudos de avaliação, revisões, consensos internacionais e propostas de estruturas conceituais, permitindo examinar aspectos relacionados ao desempenho algorítmico, equidade, segurança, qualidade dos dados e monitoramento da tecnologia.

A busca bibliográfica foi realizada na Biblioteca Virtual em Saúde (BVS), contemplando as bases Literatura Latino-Americana e do Caribe em Ciências da Saúde (LILACS) e *Medical Literature Analysis and Retrieval System Online* (MEDLINE), além da *Scientific Electronic Library Online* (SciELO) e da PubMed. A seleção dessas fontes considerou sua abrangência na área da saúde e a disponibilidade de publicações voltadas à aplicação da IA, segurança assistencial, avaliação de tecnologias e processos de tomada de decisão clínica.

Para a construção da estratégia de busca, foram utilizados descritores controlados dos vocabulários Descritores em Ciências da Saúde (DeCS) e *Medical Subject Headings* (MeSH): Inteligência Artificial (*Artificial Intelligence*), Triagem (*Triage*), Segurança do Paciente (*Patient Safety*), Equidade em Saúde (*Health Equity*) e Sistemas de Apoio à Decisão Clínica (*Clinical Decision Support Systems*). A seleção desses descritores considerou sua relação com o objetivo da pesquisa e sua capacidade de recuperar publicações pertinentes à aplicação, avaliação e segurança da IA nos processos de triagem em saúde.

Os descritores foram combinados por meio dos operadores booleanos *AND* e *OR*, respeitando as particularidades de indexação de cada base consultada. Entre as estratégias empregadas estiveram “*Artificial Intelligence*” *AND* “*Triage*”, “*Artificial Intelligence*” *AND* “*Patient Safety*”, “*Artificial Intelligence*” *AND* “*Health Equity*” e “*Artificial Intelligence*” *AND* “*Clinical Decision Support Systems*”, bem como combinações equivalentes em português nas bases que possibilitavam essa modalidade de busca. A utilização de diferentes associações permitiu abranger publicações relacionadas ao desempenho, à segurança, à equidade e ao

acompanhamento dos sistemas de IA empregados na assistência.

Foram considerados elegíveis artigos científicos disponíveis na íntegra que apresentassem relação direta com a aplicação da IA na triagem de pacientes, avaliação de desempenho algorítmico, erros de classificação, qualidade ou completude dos dados, vieses entre grupos populacionais, validação local, transparência, supervisão humana ou mecanismos de governança. Foram excluídas publicações duplicadas, editoriais, resumos sem acesso ao texto completo e materiais cujo conteúdo estivesse restrito aos aspectos computacionais do desenvolvimento de algoritmos, sem relação com o desempenho clínico, a segurança, a equidade ou a governança da tecnologia em saúde.

O processo de seleção foi conduzido inicialmente pela leitura dos títulos e resumos recuperados, seguida da avaliação integral das publicações consideradas compatíveis com a temática e com o objetivo da investigação. Após a aplicação dos critérios de elegibilidade, sete publicações científicas constituíram o corpus final da revisão, sendo selecionadas aquelas que apresentavam relação direta com o desempenho da IA na triagem, vieses algorítmicos, qualidade dos dados, avaliação local, transparência e estruturas de governança.

Para organização das informações, foram considerados autor e ano de publicação, tipo de estudo, objetivo e principais resultados, seguindo os mesmos elementos apresentados na tabela de caracterização do trabalho. Também foram observadas informações relacionadas à acurácia, subtriagem, sobretriagem, classificação inadequada, representatividade dos dados, diferenças de desempenho entre grupos populacionais, avaliação local, transparência e recomendações de acompanhamento dos sistemas, permitindo estabelecer uma análise compatível com os diferentes delineamentos presentes nas publicações selecionadas.

A análise foi realizada de forma descritiva, interpretativa e temática, buscando identificar convergências e diferenças entre as publicações quanto à acurácia da IA na triagem, ocorrência de subtriagem e sobretriagem, influência da incompletude dos dados, avaliação de desempenho entre grupos populacionais, validação local, monitoramento contínuo e responsabilização institucional. Por utilizar exclusivamente informações provenientes de publicações científicas, sem participação direta de seres humanos ou utilização de dados individuais identificáveis, não houve necessidade de submissão a Comitê de Ética em Pesquisa.

Como limitações metodológicas, reconhecem-se o caráter não sistemático da revisão narrativa, a heterogeneidade dos delineamentos incluídos, a rápida evolução das tecnologias de IA e a concentração de parte das evidências em contextos internacionais. Esses aspectos podem

restringir a generalização dos achados para diferentes serviços, populações e realidades assistenciais, especialmente diante das particularidades institucionais, da disponibilidade dos dados e das diferenças existentes entre os contextos nos quais os sistemas de IA são desenvolvidos e utilizados.

RESULTADOS E DISCUSSÃO

Na seleção final, sete publicações científicas foram incluídas e nenhuma foi excluída, pois todas apresentavam relação direta com desempenho da IA na triagem, vieses algorítmicos, qualidade dos dados, avaliação local ou mecanismos de governança. A organização do material considerou tanto pesquisas voltadas ao desempenho clínico quanto estruturas destinadas à supervisão de sistemas de IA ao longo de seu ciclo de vida, incluindo desenvolvimento, validação, implantação e monitoramento.

Tabela 1. Caracterização das publicações incluídas sobre gestão, auditoria e governança da inteligência artificial na saúde.

Autor/ano	Tipo de estudo	Objetivo	Principais resultados
Lekadir <i>et al.</i> (2025)	Consenso internacional, com abordagem Delphi modificada	Desenvolver diretrizes para o desenvolvimento e implantação de IA confiável e ética na saúde	Participaram 117 especialistas de 50 países. Foram definidos seis princípios — equidade, universalidade, rastreabilidade, usabilidade, robustez e explicabilidade — e 30 recomendações aplicáveis ao ciclo de vida da IA.
Nong <i>et al.</i> (2025)	Estudo nacional baseado no <i>American Hospital Association Annual Survey Information Technology Supplement</i> de 2023	Identificar a utilização e a avaliação local de IA e modelos preditivos nos hospitais dos Estados Unidos	A análise incluiu 2.425 hospitais; 65% utilizavam IA ou modelos preditivos, 61% avaliavam localmente a acurácia e 44% avaliavam localmente o viés.
Yi; Baik; Baek (2025)	Revisão sistemática de estudos prospectivos	Identificar os efeitos da aplicação prospectiva da IA à triagem em departamentos de emergência	Entre 1.633 publicações recuperadas, sete foram incluídas. A acurácia da triagem variou de 80,5% a 99,1%; também foram avaliadas subtriagem, sobretriagem, classificação incorreta, tempo e desfechos clínicos.
Ito <i>et al.</i> (2023)	Estudo de avaliação com vinhetas clínicas	Avaliar acurácia diagnóstica e de triagem do GPT-4 e possíveis diferenças segundo raça e etnia	Foram utilizadas 45 vinhetas e três médicos como comparação. O GPT-4 obteve 97,8% de acurácia diagnóstica e 66,7% na triagem, sem diferença estatisticamente significativa segundo raça ou etnia.
Chin <i>et al.</i> (2023)	Desenvolvimento de estrutura conceitual a partir de revisão de evidências, painel especializado e participação de partes interessadas	Formular princípios para prevenção e redução de vieses raciais e étnicos em algoritmos de saúde	Foi elaborado um modelo com cinco fases do ciclo de vida do algoritmo e cinco princípios relacionados à equidade, transparência, participação, justiça algorítmica e responsabilização.

Teeple <i>et al.</i> (2023)	Avaliação de modelo de aprendizado de máquina com dados de prontuário eletrônico	Investigar a influência de dados ausentes sobre diferenças raciais no desempenho de um modelo de triagem	Foram avaliados 61.782 atendimentos de 37.196 pacientes. A maior completude dos registros modificou o desempenho, com melhora da c-statistic entre 0,027 e 0,058 nos grupos avaliados.
Alderman <i>et al.</i> (2025)	Programa internacional de métodos mistos e consenso Delphi	Desenvolver recomendações para transparência dos conjuntos de dados e redução de vieses algorítmicos	Mais de 350 representantes de 58 países participaram da iniciativa; 194 participantes de 25 países integraram a primeira rodada Delphi. Foram estabelecidas 29 recomendações sobre documentação e uso dos dados.

Fonte: Elaborado pelos autores, 2026.

Yi, Baik e Baek (2025) encontraram acurácia da triagem entre 80,5% e 99,1%, porém o desempenho variou segundo o modelo e o nível de prioridade clínica. Ito *et al.* (2023) registraram 97,8% de acurácia diagnóstica para o GPT-4, enquanto a classificação correta da triagem foi de 66,7%. A diferença entre esses valores reforça que desempenho diagnóstico elevado não corresponde necessariamente à atribuição adequada da prioridade assistencial, sendo necessário avaliar separadamente erros de classificação capazes de antecipar ou retardar o atendimento.

Teeple *et al.* (2023) identificaram mudança no desempenho preditivo após a substituição de informações ausentes por dados mais completos, com aumento da c-statistic entre 0,027 e 0,058. Alderman *et al.* (2025) situam a composição, a representatividade e a incompletude dos conjuntos de dados entre elementos que precisam ser explicitamente documentados para reconhecer possíveis fontes de viés. A aproximação entre essas perspectivas insere a qualidade do registro clínico no processo de controle algorítmico, uma vez que desigualdades preexistentes nos dados podem modificar as previsões geradas para diferentes grupos populacionais.

Chin *et al.* (2023) organizaram a prevenção do viés em cinco fases: formulação do problema, seleção e gerenciamento dos dados, desenvolvimento e validação, implantação e integração, seguidas pelo monitoramento, manutenção, atualização ou descontinuação. A disposição sequencial dessas fases amplia a governança para além da avaliação técnica realizada antes do uso clínico, porque o comportamento do algoritmo continua sujeito a alterações depois da implantação. Na triagem, o acompanhamento posterior permite reconhecer modificações no desempenho relacionadas ao contexto institucional, à população atendida e à própria forma de utilização do sistema.

Nong *et al.* (2025) verificaram que 65% dos hospitais participantes utilizavam IA ou modelos preditivos integrados ao prontuário eletrônico, mas somente 61% avaliavam localmente

a acurácia e 44% realizavam avaliação local de viés. A diferença entre utilização e verificação do desempenho revela que a incorporação tecnológica pode ocorrer sem controle proporcional das condições em que o modelo será empregado. Esse aspecto é relevante porque variações entre serviços, populações e formas de registro podem comprometer a validade de ferramentas originalmente desenvolvidas em ambientes distintos.

Lekadir *et al.* (2025) propõem controle da qualidade das entradas e saídas, monitoramento contínuo do desempenho e auditorias periódicas capazes de verificar acurácia, consistência, equidade e segurança dos dados. A estrutura também prevê identificação de mudanças na distribuição das informações ou nas relações entre variáveis e desfechos, situação compatível com deterioração do desempenho após a implantação. A definição de responsáveis pelas auditorias e de medidas posteriores, como atualização do modelo, permite que falhas sejam tratadas por procedimento previamente estabelecido, e não somente após a ocorrência de consequências clínicas.

Yi, Baik e Baek (2025) atribuem à IA uma função de suporte à triagem realizada por profissionais, sem estabelecer evidência suficiente para substituir a participação humana. Além da variação entre algoritmos, foram considerados subtriagem, sobretriagem, fatores relacionados à classificação inadequada, tempo de atendimento e repercussões sobre o cuidado. A supervisão profissional permanece necessária porque a decisão de prioridade envolve elementos clínicos e relacionais que podem não estar integralmente representados nas variáveis processadas pelo modelo, especialmente diante de situações complexas ou informações incompletas.

Ito *et al.* (2023) não identificaram diferença estatisticamente significativa na acurácia do GPT-4 após a inclusão das categorias racial e étnica nas vinhetas clínicas, porém o experimento foi realizado com apenas 45 situações padronizadas. A própria pesquisa reconhece que esses cenários não reproduzem integralmente a complexidade da prática e que o comportamento do sistema diante de vieses raciais e étnicos no mundo real permanece incerto. A avaliação controlada, portanto, não é suficiente para dispensar verificações posteriores entre grupos populacionais quando a tecnologia for utilizada em ambientes assistenciais reais.

Teeple *et al.* (2023) trabalharam com 61.782 atendimentos referentes a 37.196 pacientes, dos quais 25.243 foram categorizados como negros e 11.953 como brancos não hispânicos. A atualização das listas de problemas modificou significativamente a c-statistic nos três modelos avaliados para ambos os grupos, embora não tenha produzido diferença racial significativa nessa métrica. Pequenas diferenças foram identificadas em outros indicadores, incluindo taxas de falsos

positivos e falsos negativos, mostrando que a auditoria de equidade não deve depender de uma única medida de desempenho.

Alderman *et al.* (2025) estabeleceram 29 recomendações divididas entre documentação dos conjuntos de dados e orientações para sua utilização, com atenção à composição, limitações e possíveis repercussões sobre diferentes grupos. A proposta inclui registro de avaliações algorítmicas e de impacto, além de medidas voltadas ao reconhecimento e à redução de riscos durante a utilização da tecnologia. Essa organização favorece a identificação de grupos sub-representados e de limitações que possam comprometer a segurança, evitando que a ausência de informações sobre os dados seja interpretada como ausência de viés.

Chin *et al.* (2023) vinculam o controle algorítmico à responsabilidade compartilhada entre desenvolvedores, usuários, instituições e agentes reguladores, com atuação nos níveis individual, organizacional e social. A equidade é acompanhada por transparência, explicabilidade, participação de pacientes e comunidades e identificação explícita de possíveis conflitos relacionados à justiça algorítmica. Para a triagem, essa estrutura permite delimitar responsabilidades sobre avaliação de desempenho, investigação de falhas, tomada de medidas corretivas e decisão sobre manutenção ou retirada de determinada ferramenta quando forem identificadas consequências desfavoráveis.

Nong *et al.* (2025) também identificaram diferenças na capacidade institucional de avaliação dos modelos: hospitais que desenvolveram suas próprias ferramentas, apresentavam maior margem operacional ou pertenciam a sistemas de saúde mostraram maior frequência de avaliação local, enquanto o uso de modelos fornecidos por terceiros esteve associado a menor probabilidade de verificação. A governança depende, portanto, de estrutura técnica e organizacional suficiente para acompanhar acurácia e viés depois da implantação. Sem essa capacidade, a expansão da IA pode ampliar diferenças entre instituições quanto à proteção oferecida aos pacientes diante de modelos inadequados ou enviesados.

CONCLUSÃO

A aplicação da IA na triagem de pacientes exige acompanhamento contínuo do desempenho algorítmico, especialmente porque a acurácia global não traduz, isoladamente, a segurança das classificações. A ocorrência de subtriagem, sobretriagem, falsos positivos e falsos negativos deve integrar a avaliação dos sistemas, juntamente com a análise da capacidade de

definir corretamente a prioridade clínica. A supervisão profissional permanece necessária diante de situações complexas, informações incompletas e possíveis limitações dos modelos.

A governança da IA deve incorporar validação local, auditorias periódicas, qualidade dos dados, transparência e avaliação do desempenho entre diferentes grupos populacionais. Registros incompletos e bases pouco representativas podem modificar a capacidade preditiva e favorecer desigualdades na assistência. A diferença entre a utilização de sistemas de IA e a frequência de avaliações locais de acurácia e viés também demonstra que a incorporação tecnológica precisa estar acompanhada de capacidade institucional para identificar alterações de desempenho e adotar medidas corretivas.

Diante da problemática proposta, a gestão e a auditoria da IA na triagem devem constituir processos permanentes, com responsabilidades definidas para monitoramento, investigação de falhas, atualização dos modelos e interrupção do uso quando houver risco à segurança. Transparência, explicabilidade, equidade, rastreabilidade e supervisão humana assumem papel central nesse processo. A contribuição do trabalho está na articulação entre desempenho técnico, qualidade dos dados, responsabilidade institucional e proteção do paciente como componentes indispensáveis à utilização segura da IA na assistência.

As limitações relacionam-se ao caráter narrativo da revisão, à heterogeneidade metodológica das publicações, à rápida evolução das tecnologias e à predominância de pesquisas realizadas em contextos internacionais. Recomenda-se que investigações futuras priorizem avaliações prospectivas e multicêntricas em serviços brasileiros, com mensuração de erros de classificação entre diferentes grupos populacionais e acompanhamento dos modelos após a implantação. Também se recomenda desenvolver e validar protocolos institucionais com indicadores objetivos de acurácia, equidade, qualidade dos dados, ocorrência de falhas e efetividade das medidas corretivas.

REFERÊNCIAS

ALDERMAN, Joseph E. *et al.* Tackling algorithmic bias and promoting transparency in health datasets: the STANDING Together consensus recommendations. **The Lancet Digital Health**, v. 7, n. 1, p. e64-e88, 2025. DOI: [https://doi.org/10.1016/S2589-7500\(24\)00224-3](https://doi.org/10.1016/S2589-7500(24)00224-3). Disponível em: <https://pubmed.ncbi.nlm.nih.gov/39701919/>.

CHIN, Marshall H. *et al.* Guiding principles to address the impact of algorithm bias on racial and ethnic disparities in health and health care. **JAMA Network Open**, v. 6, n. 12, e2345050, 2023. DOI: <https://doi.org/10.1001/jamanetworkopen.2023.45050>. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/38100101/>.

CAPPATO, Júlia Rabelo *et al.* Inteligência artificial na triagem do pronto atendimento: aplicações, benefícios e desafios em serviços de saúde. **Congresso Médico Acadêmico UniFOA**, v. 11, 2025. DOI: <https://doi.org/10.47385/cmedunifoa.2080.2025>. Disponível em: <https://conferencias.unifoa.edu.br/congresso-medvr/article/view/2080>.

DANIEL, Lucas Pedroza. Modelos de inteligência artificial na atenção primária: desempenho, transparência e segurança na triagem de pacientes. **RCMOS – Revista Científica Multidisciplinar O Saber**, v. 1, n. 1, 2023. DOI: <https://doi.org/10.51473/rcmos.v1i1.2023.1851>. Disponível em: <https://submissoesrevistarcmos.com.br/rcmos/article/view/1851>.

HARIHARAN, S. *et al.* Pediatric emergency medicine innovations using artificial intelligence for rapid clinical triage systems. **Adolescência & Saúde**, v. 21, n. 3s, 2026. DOI: <https://doi.org/10.67440/ahj.v21i3s.1004>. Disponível em: <https://adolescenciaesaude.com/index.php/aes/article/view/1004>.

HIRSCHMANN, Nelson Guilherme do Nascimento *et al.* Inovação, gestão e tecnologias em saúde: os impactos da inteligência artificial sobre os serviços. **Revista DCS**, v. 22, n. 84, 2025. DOI: <https://doi.org/10.54899/dcs.v22i84.3913>. Disponível em: <https://ojs.revistadcs.com/index.php/revista/article/view/3913>.

ITO, Naoki *et al.* The accuracy and potential racial and ethnic biases of GPT-4 in the diagnosis and triage of health conditions: evaluation study. **JMIR Medical Education**, v. 9, e47532, 2023. DOI: <https://doi.org/10.2196/47532>. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/37917120/>.

LEKADIR, Karim *et al.* FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare. **BMJ**, v. 388, e081554, 2025. DOI: <https://doi.org/10.1136/bmj-2024-081554>. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/39909534/>.

LOVADINI, Vinicius de Lima *et al.* Inteligência artificial na atenção primária brasileira: inovação assistencial, governança ética e equidade no SUS. In: **Saúde e Ciência: Estudos e Evidência. [S. l.]: Aurum Publicações**, 2026. DOI: <https://doi.org/10.63330/aurumpub.061-014>. Disponível em: <https://aurumpublicacoes.com/index.php/editora/article/view/1782>.

LOVADINI, Vinicius de Lima *et al.* Inteligência artificial na atenção primária à saúde brasileira: potencialidades, riscos éticos e implicações para a equidade no SUS. **REMUNOM**, v. 13, n. 13, 2026. DOI: <https://doi.org/10.66104/2rtma471>. Disponível em: https://remunom.ojsbr.com/multidisciplinar/pt_BR/article/view/6886.

NONG, Paige *et al.* Current use and evaluation of artificial intelligence and predictive models in US hospitals. **Health Affairs**, 2025. DOI: <https://doi.org/10.1377/hlthaff.2024.00842>. Disponível em: <https://www.healthaffairs.org/doi/10.1377/hlthaff.2024.00842>.

SILVA, Martony Demes da *et al.* Auditoria de negligência preditiva em sistemas de vigilância digital: um protocolo estagiado para mitigação de vieses em saúde pública. In: **Anais Estendidos do XXVI Simpósio Brasileiro de Computação Aplicada à Saúde. [S. l.]**:

Sociedade Brasileira de Computação, 2026. DOI:

https://doi.org/10.5753/sbcas_estendido.2026.25297. Disponível em:

https://sol.sbc.org.br/index.php/sbcas_estendido/article/view/43844.

TEEPLE, Stephanie *et al.* Exploring the impact of missingness on racial disparities in predictive performance of a machine learning model for emergency department triage. **JAMIA Open**, v. 6, n. 4, ooad107, 2023. DOI: <https://doi.org/10.1093/jamiaopen/ooad107>. Disponível em: <https://academic.oup.com/jamiaopen/article/6/4/ooad107/7484703>.

YI, Nayeon; BAIK, Dain; BAEK, Gumhee. The effects of applying artificial intelligence to triage in the emergency department: a systematic review of prospective studies. **Journal of Nursing Scholarship**, v. 57, n. 1, p. 105-118, 2025. DOI: <https://doi.org/10.1111/jnu.13024>. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/39262027/>.