O que dizem as revisões sistemáticas Cochrane sobre o uso da inteligência artificial em saúde?

Osmar Clayton Person[I], Paula Robeiro Lopes Almeida[II], Álvaro Nagib Atallah[III], Maria Eduarda dos Santos Puga[IV], Flávia Tavares Silva Elias[VI]

Faculdade Paulista de Ciências da Saúde – SPDM (FPCS), São Paulo (SP), Brasil; Universidade Federal de São Paulo (Unifesp), São Paulo (SP), Brasil; Fundação Oswaldo Cruz (Fiocruz), Brasília (DF), Brasil

RESUMO

Contextualização: A inteligência artificial (IA) surgiu com o propósito de funcionar analogamente ao cérebro humano, interpretando, analisando, descobrindo, deduzindo e relacionando informações. A Organização Mundial da Saúde (OMS) reconhece a IA como uma ferramenta para melhorar o alcance aos cuidados de saúde e a medicamentos, porém, a IA é realmente efetiva na área da saúde? **Objetivos:** Sumarizar as evidências de revisões sistemáticas realizadas pela Cochrane, referentes à precisão diagnóstica da IA em saúde. **Métodos:** Trata-se de overview de revisões sistemáticas Cochrane. Procedeu-se à busca na Cochrane Library em 2024 por meio do descritor "ARTIFICIAL INTELLIGENCE". Todas as revisões sistemáticas de estudos observacionais foram incluídas. O desfecho primário de análise foi a precisão da IA em realizar diagnósticos corretos. **Resultados:** Quatro revisões sistemáticas foram incluídas, totalizando 124 estudos primários. **Discussão:** Há poucas revisões sistemáticas realizadas pela Cochrane para avaliação da precisão diagnóstica da IA na saúde. Trata-se de estudos heterogêneos e que sugerem que a IA pode trazer benefícios. Todavia, ainda são necessários ajustes em algoritmos e melhores análises em estudos futuros que permitam melhor robustez da evidência. **Conclusão:** A IA pode ser uma ferramenta de triagem promissora na prática médica no futuro, mas, no momento, os estudos não demonstram os benefícios de seu uso. É relevante enfatizar que, devido à heterogeneidade e às limitações metodológicas dos estudos primários, o nível de evidência atual é limitado. Recomenda-se a realização de novos estudos prospectivos, com padronização das análises e relato dos resultados.

PALAVRAS-CHAVE (TERMOS DECS): Prática clínica baseada em evidências, Inteligência artificial, Estudos observacionais como assunto, Revisão sistemática.

PALAVRAS-CHAVE DOS AUTORES: Saúde, Avaliação clínica, Inovação em saúde, Precisão diagnóstica.

[I]Doutor em Saúde Baseada em Evidências pela Universidade Federal de São Paulo (Unifesp), São Paulo (SP), Brasil; Professor dos cursos de Graduação e Pós-Graduação da Faculdade Paulista de Ciências da Saúde – SPDM (FPCS), São Paulo (SP), Brasil.
https://orcid.org/0000-0002-2221-9535
[II]Doutora em Saúde Baseada em Evidências pela Universidade Federal de São Paulo (Unifesp), São Paulo (SP), Brasil.
https://orcid.org/0000-0002-4982-4831
[III]Professor Titular e Chefe da disciplina de Medicina de Urgência e Medicina Baseada em Evidências da Escola Paulista de Medicina (EPM), Universidade Federal de São Paulo (Unifesp), São Paulo (SP), Brasil; Diretor do Cochrane Brazil, São Paulo (SP), Brasil; Diretor Científico Adjunto da Associação Paulista de Medicina (APM), São Paulo (SP), Brasil.
https://orcid.org/0000-0003-0890-594X
[IV]Doutora em Saúde Baseada em Evidências pela Universidade Federal de São Paulo (Unifesp), São Paulo (SP), Brasil; Especialista em Informação no Centro Cochrane do Brasil, São Paulo (SP), Brasil.
https://orcid.org/0000-0001-8470-861X
[V]Professora e Pesquisadora em Saúde Pública da Fundação Oswaldo Cruz (Fiocruz), Brasília, Distrito Federal, Brasil; Realizou pós-doutorado na Queen's University, Kingston (ON), Canadá. Doutora em Ciências, com foco em Medicina Interna e Terapêutica e Avaliação de Tecnologias em Saúde; Professora do Mestrado Profissional em Políticas Públicas de Saúde na Fundação Oswaldo Cruz (Fiocruz), Brasília; Líder do grupo de pesquisa Evidências para Políticas e Tecnologias em Saúde.
https://orcid.org/0000-0002-7142-6266

Contribuição dos autores: Person OC: mentor, síntese de resultados e redação; Almeida PRL: extração de dados e montagem de tabelas; Atallah AN: revisão do texto e orientações; Puga MES: estratégia de busca, revisão metodológica e extração de dados; Elias FTS: revisão do texto e orientações. Todos os autores contribuíram ativamente para a discussão dos resultados do estudo e revisaram e aprovaram a versão final do trabalho para publicação.

Editor responsável por esta seção:
Álvaro Nagib Atallah. Professor Titular e Chefe da disciplina de Medicina de Urgência e Medicina Baseada em Evidências da Escola Paulista de Medicina (EPM), Universidade Federal de São Paulo (Unifesp), São Paulo (SP), Brasil; Diretor do Cochrane Brazil, São Paulo (SP), Brasil; Diretor Científico Adjunto da Associação Paulista de Medicina (APM), São Paulo (SP), Brasil.

Endereço para correspondência:
Maria Eduarda dos Santos Puga
Universidade Federal de São Paulo (Unifesp); Centro Cochrane do Brasil
Rua Sena Madureira, 1500 — Vila Clementino, São Paulo (SP), Brasil — CEP 04021-001
E-mail: mespuga@unifesp.br; mespuga@yahoo.com.br

Fonte de fomento: nenhuma. Conflito de interesse: nenhum.
Entrada: 25 de janeiro de 2025. Última modificação: 15 de abril de 2025. Aceite: 17 de abril de 2025.

CONTEXTUALIZAÇÃO

A inteligência artificial (IA) foi definida como o estudo de algoritmos que dão às máquinas a capacidade de raciocinar e executar funções como resolução de problemas, reconhecimento de objetos e palavras, inferência de estados mundiais e tomada de decisões. Embora a IA seja com frequência relacionada exclusivamente a computadores ou robôs, suas raízes são encontradas em vários campos, como a filosofia, psicologia, linguística e estatística.[1]

Os avanços na ciência da computação, como melhorias baseadas em hardware em processamento e armazenamento, permitiram a evolução das tecnologias necessárias para o advento da IA.[2]

A aplicação da IA em vários campos espalhou-se rapidamente nos últimos anos, aliada à demanda por melhorias na saúde, contribuindo para o desenvolvimento da IA no campo da medicina complementar.[3]

Vários cenários, como a crescente falta de recursos médicos, a escassez de tecnologia médica e a distribuição regional desigual de recursos médicos, contribuem para que a população deposite suas esperanças nas máquinas. Não obstante, a IA também é uma esperança para aliviar a crescente pressão sobre os hospitais e melhorar as capacidades de tratamento médico.[4]

A IA tem sido aplicada em muitos cenários da medicina, das aplicações diagnósticas em radiologia e patologia às aplicações terapêuticas e intervencionistas em cardiologia e cirurgia. Em abril de 2018, a Food and Drug Administration (FDA), dos Estados Unidos, aprovou o primeiro sistema de software que usa a IA, um programa que auxilia no diagnóstico de retinopatia diabética por meio da análise de imagens do fundo do olho. O desenvolvimento e a aplicação de tecnologias de IA em medicina tendem a crescer e é relevante que os médicos, de todas as áreas, entendam o que elas são e como essas tecnologias podem ser aproveitadas para fornecer cuidados mais seguros, eficientes e econômicos.[1]

A IA constitui um campo em rápida evolução na assistência médica, com grande potencial para informar a tomada de decisões com base em evidências e melhorar os resultados de saúde. Essa tecnologia desempenha um papel importante para suprir algumas carências dentro do sistema de saúde, como a escassez de pessoal. Além disso, pode contribuir para melhorias na administração de serviços de saúde, como cobrança, reembolsos e exposição a fraudes de seguros.[5]

A possibilidade da IA melhorar a eficiência na saúde passou também a ser objeto de discussão na sociedade.[5] Por esses motivos, buscou-se as melhores evidências disponíveis relativas à efetividade da IA na saúde. Considerando a metodologia utilizada pela Colaboração Cochrane na busca das melhores evidências em saúde, ateve-se às revisões sistemáticas realizadas pela Cochrane acerca do tópico em questão.

OBJETIVOS

O estudo visa sumarizar as evidências de revisões sistemáticas realizadas pela Cochrane referentes à precisão diagnóstica da IA em saúde.

METODOLOGIA

Desenho de estudo

Trata-se de overview de revisões sistemáticas publicadas na Cochrane Library. Não houve restrições relativas ao local, data e idioma em que os estudos foram publicados.

CRITÉRIOS DE INCLUSÃO

Tipos de participantes

Incluíram-se todas as revisões sistemáticas de estudos observacionais em humanos que utilizaram IA em saúde e que constam no banco de dados Cochrane Library. Não houve restrição de idade para inclusão dos participantes.

Tipos de intervenções

Consideraram-se todos os estudos a respeito do uso de IA em saúde. A IA foi comparada ao gold standard para o diagnóstico das condições em análise ou a qualquer controle.

Tipos de resultados

Para o desfecho primário de análise, avaliou-se a precisão da IA na saúde.

PROCESSO DE BUSCA E SELEÇÃO DE ESTUDOS

A busca por revisões sistemáticas foi realizada em 2 de dezembro de 2024 na Cochrane Library, utilizando a terminologia oficial do Medical Subject Headings (MeSH) e da Cochrane Library, via Wiley On-Line Library. Utilizou-se o descritor "ARTIFICIAL INTELLIGENCE", conforme a **Tabela 1**.

As análises dos estudos e a extração dos dados foram realizadas respeitando os critérios de inclusão descritos. Todo o

Tabela 1. Estratégia de busca

	Estratégia	Resultados
#1	MeSH descriptor: [ARTIFICIAL INTELLIGENCE] this term only	7
ID	Search Hits	
#1	MeSH descriptor: [ARTIFICIAL INTELLIGENCE] this term only	7
Date run	12/02/2024 09:15:58	

processo de extração de dados foi realizado por dois pesquisadores independentes.

As revisões encontradas foram analisadas a partir do texto completo e a extração dos dados foi realizada a partir dos arquivos originais das revisões sistemáticas.

Utilizou-se uma folha de extração predeterminada, contendo os seguintes pontos principais: ano de publicação; nome dos autores e título da revisão; número de estudos primários; tipos e número de participantes; intervenções e resultados; análise de viés e suas justificativas; detalhes de grupos de intervenção; duração e parâmetros; período de acompanhamento e, quando presentes, valores estatísticos em metanálise; risco relativo; diferenças entre médias padronizadas, ou não padronizadas, e intervalo de confiança.

As análises quantitativas utilizadas das variáveis contínuas foram agrupadas em diferença média (*mean difference*, MD) ou diferença média padronizada (*standardized mean difference*, SMD), com intervalos de confiança de 95% (95% CI).

RESULTADOS

A estratégia de busca recuperou, em dezembro de 2024, um total de sete citações na Cochrane Library. Dessas, quatro revisões sistemáticas atenderam aos critérios de inclusão deste estudo. Todos as quatro revisões sistemáticas foram incluídas, totalizando 124 estudos primários, conforme a **Tabela 2**.[6-9]

Tabela 2. Característica dos estudos incluídos

Autoria (ano)	Amostra	Objetivo	Resultado	Conclusão
Vandevenne et al. (2023)[6]	63 estudos transversais e diagnósticos de caso controle	Avaliar a precisão diagnóstica de algoritmos de IA para detectar ceratocone em pessoas que apresentam erros de refração, especialmente aquelas cuja visão não pode mais ser totalmente corrigida com óculos e aquelas que buscam cirurgia refrativa da córnea e aquelas com suspeita de ceratocone. A IA pode ajudar oftalmologistas, optometristas e outros profissionais de cuidados oculares a tomar decisões sobre encaminhamento para especialistas em córnea.	Incluíram-se 63 estudos, publicados entre 1994 e 2022, que desenvolveram e investigaram a precisão da IA para o diagnóstico de ceratocone. Havia três unidades diferentes de análise nos estudos: olhos, participantes e imagens. Quarenta e quatro analisaram 23.771 olhos, quatro estudos analisaram 3.843 participantes e quinze estudos analisaram 38.832 imagens. Cinquenta e quatro artigos avaliaram a detecção de ceratocone manifesto, definido como uma córnea que mostrou qualquer sinal clínico de ceratocone. A precisão da IA parece quase perfeita, com uma sensibilidade resumida de 98,6% (95% CI – 97,6% a 99,1%) e uma especificidade resumida de 98,3% (95% CI – 97,4% a 98,9%). No entanto, a precisão variou entre os estudos e a certeza da evidência foi baixa. Vinte e oito artigos avaliaram a detecção de ceratocone subclínico, embora a definição de subclínico tenha variado. Foram agrupados ceratocone subclínico e olhos muito assimétricos. Os testes mostraram boa precisão, com uma sensibilidade resumida de 90% (95% CI – 84,5% a 93,8%) e uma especificidade resumida de 95,5% (95% CI – 91,9% a 97,5%). Contudo, a certeza da evidência foi muito baixa para sensibilidade e baixa para especificidade. Em ambos os grupos, a maioria dos estudos foi classificada como de alto risco de viés, com altas preocupações de aplicabilidade, no domínio da seleção de pacientes, já que a maioria era composta de estudos de caso-controle. Além disso, a certeza da evidência foi baixa a muito baixa devido ao viés de seleção, inconsistência e imprecisão. Não foi possível explicar a heterogeneidade entre os estudos. As análises de sensibilidade baseadas no desenho do estudo, algoritmo de IA, técnica de imagem (topografia versus tomografia) e fonte de dados (parâmetros versus imagens) não mostraram diferenças nos resultados.	A IA parece ser uma ferramenta de triagem promissora na prática oftalmológica para o diagnóstico de ceratocone. A precisão do teste foi muito alta para ceratocone manifesto e ligeiramente menor para ceratocone subclínico, indicando uma maior chance de perder um diagnóstico em pessoas sem sinais clínicos. Isso pode levar à progressão do ceratocone ou à indicação errônea para cirurgia refrativa, o que pioraria a doença. Não há como ter conclusões claras e confiáveis devido ao alto risco de viés, à heterogeneidade inexplicada dos resultados e às altas preocupações com a aplicabilidade, o que reduziu a confiança nas evidências. Uma maior padronização em pesquisas futuras aumentaria a qualidade dos estudos e melhoraria a comparabilidade entre eles.

Continua...

Tabela 2. Continuação.

Autoria (ano)	Amostra	Objetivo	Resultado	Conclusão
Kang et al. (2024)[7]	36 estudos transversais	Avaliar a precisão diagnóstica da IA como ferramenta de triagem para degeneração macular (DM) relacionada à idade.	Foram identificados 36 estudos elegíveis que relataram quarenta conjuntos de dados de desempenho de algoritmos, abrangendo mais de 16 mil participantes e 62 mil imagens. Foram incluídos 28 estudos (78%) que relataram 31 algoritmos com dados de desempenho na meta-análise. Os estudos restantes (25%) relataram oito algoritmos –apresentados na síntese qualitativa – que não tinham dados de desempenho utilizáveis.A maioria dos estudos foi conduzida na Ásia, seguida pela Europa, Estados Unidos e esforços colaborativos abrangendo vários países. A maioria dos estudos envolveu participantes em ambiente hospitalar, enquanto outros usaram imagens da retina de repositórios públicos. Alguns estudos não especificaram fontes de imagem. Com base em quatro dos 36 estudos que relataram informações demográficas, a idade dos participantes do estudo variou de 62 a 82 anos. Os algoritmos incluídos usaram vários tipos de imagens da retina como entrada do modelo, como imagens de tomografia de coerência óptica (OCT) (n = 15), imagens do fundo (n = 6) e imagens multimodais (n = 7). O método principal predominante usado foi de redes neurais profundas. Todos os estudos que relataram algoritmos validados externamente estavam com alto risco de viés, principalmente devido ao potencial viés de seleção de um desenho de duas portas ou à exclusão inadequada de imagens da retina potencialmente elegíveis (ou participantes). Apenas três dos quarenta algoritmos incluídos foram validados externamente (7,5%, 3/40). A sensibilidade e especificidade resumidas foram 0,94 (95% CI – 0,9 a 0,97) e 0,99 (95% CI – 0,76 a 1), respectivamente, quando comparados a classificadores humanos (três estudos; 27.872 imagens; evidência de baixa certeza). A prevalência de imagens com DM variou de 0,3% a 49%. Vinte e oito algoritmos foram validados internamente (20%, 8/40) ou testados em um conjunto de desenvolvimento (50%, 20/40); a sensibilidade e a especificidade combinadas foram de 0,93 (95% CI – 0,89 a 0,96) e 0,96 (95% CI – 0,94 a 0,98), respectivamente, quando comparados a classificadores humanos (28 estudos; 33.409 imagens; evidência de baixa certeza). Não foram identificadas fontes significativas de heterogeneidade entre esses 28 algoritmos. Embora os algoritmos que usam imagens de OCT parecessem mais homogêneos e tivessem a maior especificidade de resumo (0,97, 95% CI – 0,93 a 0,98), eles não foram superiores aos algoritmos que usam imagens de fundo sozinhas (0,94, 95% CI 0,89 a 0,97) ou imagens multimodais (0,96, 95% CI – 0,88 a 0,99; p para meta-regressão = 0,239). A prevalência mediana de imagens foi de 30% (intervalo interquartil [IQR] 22% a 39%). Não foram incluídos oito estudos que descreveram nove algoritmos (um estudo relatou dois conjuntos de resultados de algoritmos) para distinguir DM de imagens normais, imagens de outras doenças ou outras lesões retinianas não relacionadas na meta-análise. Cinco desses algoritmos foram geralmente baseados em conjuntos de dados menores (intervalo de 21 a 218 participantes por estudo), mas com uma prevalência maior de imagens de DM (IC de 33% a 66%). Em relação aos classificadores humanos, a sensibilidade relatada nesses estudos variou de 0,95 e 0,97, enquanto a especificidade variou de 0,94 a 0,99. Da mesma forma, usaram-se pequenos conjuntos de dados (intervalo de 46 a 106), quatro algoritmos adicionais para detectar DM de outras lesões retinianas mostraram alta sensibilidade (intervalo de 0,96 a 1) e especificidade (intervalo de 0,77 a 1).	Evidências de baixa a muito baixa certeza sugerem que um teste baseado em algoritmo pode identificar corretamente a maioria dos indivíduos com DM, sem aumentar encaminhamentos desnecessários (falsos positivos) em ambientes de cuidados primários ou especializados. Houve preocupações significativas para aplicar os resultados da revisão devido a variações na prevalência de DM nos estudos incluídos. Além disso, entre os testes baseados em algoritmos, as estimativas de precisão diagnóstica estavam com risco de viés devido aos participantes do estudo não refletirem características do mundo real, validação inadequada do modelo e a probabilidade de relatórios de resultados seletivos. A qualidade e a quantidade limitadas de algoritmos validados externamente destacaram a necessidade de evidências de alta certeza. Essas evidências exigirão uma definição padronizada para DM em diferentes modalidades de imagem e validação externa do algoritmo para avaliar a generalização.
Chuchu et al. (2018)[8]	2 coortes	Avaliar a precisão diagnóstica de aplicativos de smartphones para descartar melanoma invasivo cutâneo e variantes melanocíticas intraepidérmicas atípicas em adultos com preocupações sobre lesões cutâneas suspeitas.	Foram incluídas duas coortes. Ambos os estudos apresentaram alto risco de viés devido ao recrutamento seletivo de participantes e altas taxas de imagens não avaliáveis. As preocupações sobre a aplicabilidade dos achados foram altas devido à inclusão apenas de lesões já selecionadas para excisão em um ambiente de clínica dermatológica e à aquisição de imagens por clínicos em vez de usuários de aplicativos de smartphone. Relataram-se dados para cinco aplicativos de celular e 332 lesões cutâneas suspeitas com 86 melanomas nos dois estudos. Nos quatro aplicativos baseados em IA que classificaram imagens de lesões (fotografias) como melanomas (um aplicativo) ou como lesões de alto risco ou "problemáticas" (três aplicativos) usando um algoritmo pré-programado, as sensibilidades variaram de 7% (95% CI – 2% a 16%) a 73% (95% CI – 52% a 88%) e as especificidades de 37% (95% CI – 29% a 46%) a 94% (95% CI – 87% a 97%). O único aplicativo usando revisão de armazenar e encaminhar imagens de lesões por um dermatologista teve uma sensibilidade de 98% (95% CI – 90% a 100%) e especificidade de 30% (95% CI – 22% a 40%). O número de falhas de teste (imagens de lesões analisadas pelos aplicativos, mas classificadas como "não avaliáveis" e excluídas pelos autores do estudo) variou de três a 31 (ou 2% a 18% das lesões analisadas). O aplicativo store-and-forward apresentou uma das maiores taxas de falha de teste (15%). Pelo menos um melanoma foi classificado como não avaliável em três das quatro avaliações do aplicativo.	Os aplicativos de smartphone que usam análise baseada em IA ainda não demonstraram promessa suficiente em termos de precisão e estão associados a uma alta probabilidade de não detectar melanomas. Os aplicativos baseados em imagens de armazenamento e encaminhamento podem ter um papel potencial na apresentação oportuna de pessoas com lesões potencialmente malignas, facilitando práticas ativas de autogerenciamento de saúde e engajamento precoce daqueles com lesões cutâneas suspeitas. Entretanto, eles podem incorrer em um aumento significativo de recursos e carga de trabalho. Dada a escassez de evidências e a baixa qualidade metodológica dos estudos existentes, não é possível concluir sobre essa prática. Contudo, esse é um campo que avança rapidamente, e novos e melhores aplicativos com relatórios robustos de estudos podem mudar essas conclusões substancialmente.

Continua...

Tabela 2. Continuação.

Autoria (ano)	Amostra	Objetivo	Resultado	Conclusão
Ferrante di Ruffano et al. (2018)[9]	23 coortes	Avaliar a precisão dos sistemas diagnósticos assistidos por computador (SDC) para o diagnóstico de melanoma invasivo cutâneo e variantes melanocíticas intraepidérmicas atípicas, carcinoma basocelular e carcinoma espinocelular em adultos, e comparar sua precisão com a da dermatoscopia.	Foram incluídos 42 estudos, 24 avaliando sistemas SDC baseados em dermatoscopia digital (Derm-CAD) em 23 coortes com 9.602 lesões (1.220 melanomas, pelo menos 83 CBC, 9 CEC), fornecendo 32 conjuntos de dados para Derm-CAD e sete para dermatoscopia. Dezoito estudos avaliaram SDC baseado em espectroscopia (Spectro-SDC) em dezesseis coortes de estudo com 6.336 lesões (934 melanomas, 163 CBC, 49 CEC), fornecendo 32 conjuntos de dados para Spectro-SDC e seis para dermatoscopia. Estes consistiram em quinze estudos usando imagens multiespectrais (MSI), dois estudos usando espectroscopia de impedância elétrica (EIS) e um estudo usando espectroscopia de refletância difusa. Os estudos foram relatados de forma incompleta e com risco de viés pouco claro a alto em todos os domínios. Os estudos incluídos abordam inadequadamente a questão da revisão devido à abundância de estudos de baixa qualidade, aos relatórios precários e ao recrutamento de grupos altamente selecionados de participantes. Encontrou-se, em todos os sistemas SDC, uma variação considerável nas tecnologias de hardware e software usadas, nos tipos de algoritmo de classificação empregados, nos métodos usados para treinar os algoritmos e em quais características morfológicas da lesão foram extraídas e analisadas em todos os sistemas SDC, e até mesmo entre estudos que avaliaram sistemas SDC. A metanálise mostrou que os sistemas SDC tinham alta sensibilidade para a identificação correta de melanoma invasivo cutâneo e variantes melanocíticas intraepidérmicas atípicas em populações altamente selecionadas, mas com especificidade baixa e muito variável, particularmente para sistemas Spectro-SDC. Os dados agrupados de 22 estudos estimaram a sensibilidade do Derm-SDC para a detecção de melanoma em 90,1% (95% CI – 84,0% a 94,0%) e especificidade em 74,3% (95% CI – 63,6% a 82,7%). Os dados agrupados de oito estudos estimaram a sensibilidade da SDC de imagem multiespectral (MSI-SDC) em 92,9% (95% CI – 83,7% a 97,1%) e a especificidade em 43,6% (95% CI – 24,8% a 64,5%). Quando aplicado a uma população hipotética de mil lesões na prevalência média observada de melanoma de 20%, o Derm-SDC deixaria de detectar vinte melanomas e levaria a 206 resultados falso-positivos para melanoma. O MSI-SDC deixaria de detectar catorze melanomas e levaria a 451 diagnósticos falsos para melanoma. As descobertas preliminares sugerem que os sistemas SDC são pelo menos tão sensíveis quanto a avaliação de imagens dermatoscópicas para o diagnóstico de melanoma invasivo e variantes melanocíticas intraepidérmicas atípicas. Entretanto, não é possível fazer considerações agrupadas sobre o uso de SDC em populações não encaminhadas, ou sua precisão na detecção de neoplasias queratinocíticas, ou seu uso em qualquer cenário como auxílio diagnóstico, devido à escassez de estudos.	Em populações de pacientes altamente selecionadas, todos os tipos de SDC demonstram alta sensibilidade e podem ser úteis como um backup para diagnóstico especializado para auxiliar na minimização do risco de melanomas perdidos. No entanto, a base de evidências é atualmente muito pobre para entender se as saídas do SDC se traduzem em diferentes tomadas de decisão clínicas na prática. Há dados insuficientes sobre o uso de SDC em ambientes comunitários ou para a detecção de neoplasias queratinocíticas. A base de evidências para sistemas individuais é muito limitada para conclusões sobre quais podem ser preferidos para a prática. Estudos comparativos prospectivos são necessários para avaliar o uso de SDC como auxiliares de diagnóstico, em comparação com a dermatoscopia face a face e em populações participantes que sejam representativas daquelas nas quais o teste seria usado na prática.

95% CI = intervalo de confiança de 95%; CBC: carcinoma basocelular; CEC: carcinoma espinocelular.

DISCUSSÃO

A IA corresponde a uma subárea da ciência da computação que se dedica à pesquisa e à proposição de dispositivos computacionais capazes de simular alguns aspectos do intelecto humano, como a capacidade de raciocínio, percepção, tomada de decisões e resolução de problemas.[10]

Nos últimos anos, ampliou-se o uso da IA em medicina e o debate social a respeito dessa ferramenta. O uso de computadores que podem analisar um grande volume de dados com

base em algoritmos definidos por especialistas, em tese, podem ser capazes de propor soluções para problemas médicos.[11]

Não obstante, a IA é cada vez mais empregada em áreas de saúde, como oncologia, radiologia e dermatologia,[12] mas há questões que, todavia, não podem passar despercebidas diante da realidade da IA na saúde. Uma das mais importantes é a ética, sendo imperativo estabelecer diretrizes que garantam transparência e confiabilidade nas decisões médicas assistidas por IA.[13]

O estudo avaliou a precisão diagnóstica da IA na área da saúde nas revisões sistemáticas de pesquisas observacionais desenvolvidas pela Colaboração Cochrane. Nesse âmbito, encontraram-se quatro publicações, as quais totalizaram 124 estudos primários.

Vandevenne et al.[6] avaliaram a precisão diagnóstica de algoritmos de IA para detectar ceratocone em pessoas que apresentavam erros de refração, especialmente aquelas cuja visão não podia mais ser totalmente corrigida com óculos, aquelas que buscam cirurgia refrativa da córnea e aquelas com suspeita de ceratocone. Incluíram-se 63 estudos observacionais, publicados entre 1994 e 2022, que desenvolveram e investigaram a precisão da IA para o diagnóstico de ceratocone. Havia três unidades diferentes de análise nos estudos: olhos, participantes e imagens. Quarenta e quatro estudos analisaram 23.771 olhos, 4 analisaram 3.843 participantes, e 15 analisaram 38.832 imagens. Cinquenta e quatro artigos avaliaram a detecção de ceratocone manifesto, definido como uma córnea que mostrou qualquer sinal clínico de ceratocone. A precisão da IA pareceu quase perfeita, com uma sensibilidade resumida de 98,6% (95% CI – 97,6% a 99,1%) e uma especificidade resumida de 98,3% (95% CI – 97,4% a 98,9%). No entanto, a precisão variou entre os estudos e a certeza da evidência foi baixa.

Vinte e oito artigos avaliaram a detecção de ceratocone subclínico, embora a definição de subclínico tenha variado nos estudos. Foram agrupados ceratocone subclínico e olhos muito assimétricos. Os testes mostraram precisão, com uma sensibilidade resumida de 90,0% (95% CI – 84,5% a 93,8%) e uma especificidade resumida de 95,5% (95% CI – 91,9% a 97,5%). Entretanto, a certeza da evidência foi baixa para a sensibilidade e para a especificidade.

Em ambos os grupos, a maioria dos estudos foi classificada como de alto risco de viés, com altas preocupações de aplicabilidade, no domínio da seleção de pacientes, já que a maioria era composta de estudos de caso-controle. Além disso, a certeza da evidência foi baixa e muito baixa devido ao viés de seleção, inconsistência e imprecisão.

Não foi possível explicar a heterogeneidade entre os estudos. As análises de sensibilidade baseadas no desenho do estudo, algoritmo de IA, técnica de imagem (topografia versus tomografia) e fonte de dados (parâmetros versus imagens) não mostraram diferenças nos resultados.

Vandevenne et al.[6] consideraram que a IA parece ser uma ferramenta de triagem promissora na prática oftalmológica para o diagnóstico de ceratocone. A precisão do teste foi alta para ceratocone manifesto e ligeiramente menor para ceratocone subclínico, indicando uma maior chance de perder um diagnóstico em pessoas sem sinais clínicos. Isso pode levar à progressão do ceratocone ou a uma indicação errônea para cirurgia refrativa, o que pioraria a doença. Nesse contexto, não é possível obter conclusões claras e confiáveis devido ao alto risco de viés, à heterogeneidade inexplicada dos resultados e às preocupações com a aplicabilidade. Tudo isso reduziu a confiança nas evidências. Uma maior padronização em pesquisas futuras aumentaria a qualidade dos estudos e melhoraria a comparabilidade entre eles.

Kang et al.[7] realizaram uma revisão sistemática para avaliar a precisão diagnóstica da IA como ferramenta de triagem para degeneração macular (DM) relacionada à idade. Foram incluídos 36 estudos observacionais que relataram quarenta conjuntos de dados de desempenho de algoritmos, abrangendo mais de 16 mil participantes e 62 mil imagens. Vinte e oito estudos (78%) relataram 31 algoritmos com dados de desempenho na metanálise. Os estudos restantes (25%) relataram oito algoritmos que não tinham dados de desempenho utilizáveis; estes foram relatados na síntese qualitativa. Os algoritmos incluídos usaram vários tipos de imagens da retina como entrada do modelo, como imagens de tomografia de coerência óptica (OCT) (n = 15), imagens do fundo (n = 6) e imagens multimodais (n = 7). O método principal foi de redes neurais profundas. Todos os estudos que relataram algoritmos validados externamente estavam com alto risco de viés, sobretudo devido ao potencial viés de seleção de um desenho de duas portas ou à exclusão inadequada de imagens da retina potencialmente elegíveis, ou participantes.

Apenas três dos quarenta algoritmos incluídos foram validados externamente (7,5%, 3/40). A sensibilidade e especificidade resumidas foram 0,94 (95% CI – 0,90 a 0,97) e 0,99 (95% CI – 0,76 a 1), respectivamente, quando comparados a classificadores humanos (3 estudos; 27.872 imagens; evidência de baixa certeza). A prevalência de imagens com DM variou de 0,3% a 49%. Vinte e oito algoritmos foram validados internamente (20%, 8/40) ou testados em um conjunto de desenvolvimento (50%, 20/40); a sensibilidade e a especificidade combinadas foram de 0,93 (95% CI – 0,89 a 0,96) e 0,96 (95% CI – 0,94 a 0,98), respectivamente, quando comparados a classificadores humanos (28 estudos; 33.409 imagens; evidência de baixa certeza). Não foram identificadas fontes significativas de heterogeneidade entre esses 28 algoritmos. Embora os algoritmos que usam imagens de OCT parecessem mais homogêneos e tivessem a maior especificidade de resumo (0,97, 95% CI – 0,93 a 0,98), eles não foram superiores aos algoritmos que usam imagens de fundo sozinhas (0,94, 95% CI 0,89 a 0,97) ou imagens multimodais (0,96, 95% CI – 0,88 a 0,99; p para metarregressão = 0,239).

A prevalência mediana de imagens foi de 30% (intervalo interquartil [IQR] 22% a 39%).

Não foram incluídos oito estudos que descreveram nove algoritmos (um estudo relatou dois conjuntos de resultados de algoritmos) para distinguir DM de imagens normais, imagens de outras doenças ou outras lesões retinianas não relacionadas na meta-análise. Cinco desses algoritmos foram geralmente baseados em conjuntos de dados menores (intervalo de 21 a 218 participantes por estudo), mas com uma prevalência maior de imagens de DM (IC de 33% a 66%). Em relação aos classificadores humanos, a sensibilidade relatada nesses estudos variou de 0,95 e 0,97, enquanto a especificidade variou de 0,94 a 0,99. Usaram-se também pequenos conjuntos de dados (intervalo de 46 a 106), quatro algoritmos adicionais para detectar DM de outras lesões retinianas mostraram alta sensibilidade (intervalo de 0,96 a 1) e especificidade (intervalo de 0,77 a 1).

Kang et al.[7] concluíram que evidências de baixa a muito baixa certeza sugerem que um teste baseado em algoritmo pode identificar corretamente a maioria dos indivíduos com DM, sem aumentar encaminhamentos desnecessários (falsos positivos) em ambientes de cuidados primários ou especializados. Entretanto, há preocupações significativas para aplicar os resultados da revisão devido a variações na prevalência de DM nos estudos incluídos. Além disso, entre os testes baseados em algoritmos, as estimativas de precisão diagnóstica estavam com risco de viés devido aos participantes do estudo não refletirem características do mundo real, validação inadequada do modelo e a probabilidade de relatórios de resultados seletivos. A qualidade e a quantidade limitadas de algoritmos validados externamente destacaram a necessidade de evidências de alta certeza. Essas evidências exigirão uma definição padronizada para DM em diferentes modalidades de imagem e validação externa do algoritmo para avaliar a generalização.

A revisão sistemática realizada por Chuchu et al.[8] envolveu dois estudos coorte, que avaliaram a precisão diagnóstica de aplicativos de smartphones para descartar o melanoma invasivo cutâneo e as variantes melanocíticas intraepidérmicas atípicas em adultos com lesões cutâneas suspeitas.

Ambos os estudos apresentaram alto risco de viés devido ao recrutamento seletivo de participantes e altas taxas de imagens não avaliáveis. As preocupações sobre a aplicabilidade dos achados foram altas devido à inclusão apenas de lesões já selecionadas para excisão em um ambiente de clínica dermatológica e à aquisição de imagens por clínicos em vez de usuários de aplicativos de smartphone. Relataram-se dados para cinco aplicativos de celular e 332 lesões cutâneas suspeitas com 86 melanomas nos dois estudos. Nos quatro aplicativos baseados em IA que classificaram imagens de lesões como melanomas (um aplicativo) ou como lesões de alto risco ou "suspeitas" (três aplicativos) usando um algoritmo pré-programado, as sensibilidades variaram de 7% (95% CI – 2% a 16%) a 73% (95% CI – 52% a 88%) e as especificidades de 37% (95% CI – 29% a 46%) a 94% (95% CI – 87% a 97%). O único aplicativo que usou armazenamento e encaminhou as imagens de lesão por um dermatologista teve uma sensibilidade de 98% (95% CI – 90% a 100%) e especificidade de 30% (95% CI – 22% a 40%).

O número de falhas de teste (imagens de lesões analisadas pelos aplicativos, mas classificadas como "não avaliáveis" e excluídas pelos autores do estudo) variou de três a 31 (ou 2% a 18% das lesões analisadas). O aplicativo store-and-forward apresentou uma das maiores taxas de falha de teste (15%). Pelo menos um melanoma foi classificado como não avaliável em três das quatro avaliações do aplicativo.

Chuchu et al.[8] concluíram que aplicativos de smartphone que usam análise baseada em IA ainda não demonstraram resultados suficientes em termos de precisão e estão associados à alta probabilidade de não detectar melanomas. Os aplicativos baseados em imagens de armazenamento e encaminhamento podem ter um papel potencial na apresentação oportuna de pessoas com lesões potencialmente malignas, facilitando práticas ativas de autogerenciamento de saúde e engajamento precoce daqueles com lesões cutâneas suspeitas. Entretanto, eles podem incorrer em um aumento significativo de recursos e carga de trabalho. No contexto da escassez de evidências e da baixa qualidade metodológica dos estudos disponíveis até o momento, não é possível concluir a respeito dessa prática como rotina. Todavia, esse é um campo que avança rapidamente, e novos e melhores aplicativos com relatórios robustos de estudos podem mudar essas conclusões.

Em outra revisão sistemática, Ferrante di Ruffano et al.[9] avaliaram a precisão dos sistemas diagnósticos assistidos por computador (SDC) para o diagnóstico de melanoma invasivo cutâneo e variantes melanocíticas intraepidérmicas atípicas, carcinoma basocelular e carcinoma espinocelular em adultos, e compararam sua precisão com a da dermatoscopia. Incluíram-se 23 coortes com 9.602 lesões (1.220 melanomas, pelo menos 83 carcinomas basocelulares [CBC] e nove carcinomas espinocelulares [CEC]), fornecendo 32 conjuntos de dados para o sistema digital Derm-CAD e sete para a dermatoscopia. Dezoito estudos avaliaram SDC baseado em espectroscopia (Spectro-SDC) em dezesseis coortes de estudo com 6.336 lesões (934 melanomas, 163 CBC, 49 CEC), fornecendo 32 conjuntos de dados para Spectro-SDC e seis para dermatoscopia. Estes consistiram em quinze estudos usando imagens multiespectrais (MSI), dois estudos usando espectroscopia de impedância elétrica (EIS) e um estudo usando espectroscopia de refletância difusa. Os estudos foram relatados de forma incompleta e com risco de viés incerto a alto em todos os domínios. Os estudos incluídos abordam inadequadamente a questão da revisão devido à abundância de estudos de baixa

qualidade, aos relatórios precários e ao recrutamento de grupos altamente selecionados de participantes.

Em todos os sistemas SDC, encontrou-se uma variação considerável nas tecnologias de hardware e software usadas, nos tipos de algoritmo de classificação empregados, nos métodos usados para treinar os algoritmos e em quais características morfológicas da lesão foram extraídas e analisadas em todos os sistemas SDC, e até mesmo entre estudos que avaliaram sistemas SDC. A metanálise mostrou que os sistemas SDC tinham alta sensibilidade para a identificação correta de melanoma invasivo cutâneo e variantes melanocíticas intraepidérmicas atípicas em populações altamente selecionadas, mas com especificidade baixa e muito variável, particularmente para os sistemas Spectro-SDC. Os dados agrupados de 22 estudos estimaram a sensibilidade do Derm-SDC para a detecção de melanoma em 90,1% (95% CI – 84,0% a 94,0%) e a especificidade em 74,3% (95% CI – 63,6% a 82,7%). Os dados agrupados de oito estudos estimaram a sensibilidade da SDC de imagem multiespectral (MSI-SDC) em 92,9% (95% CI – 83,7% a 97,1%) e a especificidade em 43,6% (95% CI – 24,8% a 64,5%). Quando aplicado a uma população hipotética de mil lesões na prevalência média observada de melanoma de 20%, o Derm-SDC deixaria de detectar vinte melanomas e levaria a 206 resultados falso-positivos para melanoma. O MSI-SDC deixaria de detectar catorze melanomas e levaria a 451 diagnósticos falsos para melanoma. As descobertas preliminares sugerem que os sistemas SDC são pelo menos tão sensíveis quanto à avaliação de imagens dermatoscópicas para o diagnóstico de melanoma invasivo e variantes melanocíticas intraepidérmicas atípicas. Entretanto, não é possível fazer considerações agrupadas sobre o uso de SDC em populações não encaminhadas, ou sua precisão na detecção de neoplasias queratinocíticas, ou seu uso em qualquer cenário como auxílio diagnóstico, devido à escassez de estudos.

Ferrante di Ruffano et al.[9] concluíram que, em populações de pacientes altamente selecionadas, todos os tipos de SDC demonstram alta sensibilidade e podem ser úteis como um backup para diagnóstico especializado para auxiliar na minimização do risco de melanomas não diagnosticados. No entanto, a base de evidências é atualmente muito pobre para entender se as saídas do SDC se traduzem em diferentes tomadas de decisão clínicas na prática. Há dados insuficientes sobre o uso de SDC em ambientes comunitários ou para a detecção de neoplasias queratinocíticas. A base de evidências para sistemas individuais é muito limitada para conclusões sobre quais podem ser preferidos para a prática. São necessários, portanto, estudos comparativos prospectivos para avaliar o uso de SDC como auxiliares de diagnóstico, em comparação com a dermatoscopia face a face e em populações participantes que sejam representativas daquelas nas quais o teste seria usado na prática.

Diante dos achados dos estudos incluídos nesta revisão, a precisão diagnóstica da IA na área da saúde ainda não está plenamente estabelecida. A IA poderá ser uma ferramenta de triagem promissora na medicina do futuro, mas, para melhorar a performance, deverão ocorrer ajustes nos algoritmos atuais. Nesse âmbito, também serão necessários novos estudos prospectivos para melhor robustez da evidência do uso da IA na saúde. Esses estudos deverão ter padronização de relato de resultados para facilitar a busca por evidência de alta qualidade. Até lá, a IA permanece como uma promessa de melhora da precisão diagnóstica na área da saúde.

CONCLUSÃO

Há poucas revisões sistemáticas Cochrane que avaliaram a precisão diagnóstica da IA na área da saúde. Embora a IA possa ser uma ferramenta de triagem promissora na prática médica no futuro, no momento, a maioria dos estudos não demonstra benefícios consideráveis em seu uso. É relevante enfatizar que o nível de evidência atual é bastante limitado devido à heterogeneidade e às limitações metodológicas dos estudos primários. Nesse contexto, recomenda-se a realização de novos estudos prospectivos, com padronização das análises e relato dos resultados.

REFERÊNCIAS

1. Hashimoto DA, Witkowski E, Gao L, Meireles O, Rosman G. Artificial intelligence in anesthesiology: current techniques, clinical applications, and limitations. Anesthesiology. 2020 Feb;132(2):379–94. PMID: 31939856; https://doi.org/10.1097/ALN.0000000000002960.
2. Buchanan BG. A (very) brief history of artificial intelligence. AIMag [Internet]. 2005 Dec 15 [cited 2025 Apr 14];26(4):53. Available from: https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/1848.
3. Mekki YM, Zughaier SM. Teaching artificial intelligence in medicine. Nat Rev Bioeng. 2024;2:450–1. https://doi.org/10.1038/s44222-024-00195-0.
4. Gou F, Liu J, Xiao C, Wu J. Research on artificial-intelligence-assisted medicine: a survey on medical artificial intelligence. Diagnostics (Basel). 2024 Jul 9;14(14):1472. PMID: 39061610; https://doi.org/10.3390/diagnostics14141472.
5. Okeibunor JC, Jaca A, Iwu-Jaja CJ, et al. The use of artificial intelligence for delivery of essential health services across WHO regions: a scoping review. Front Public Health. 2023 Jul 4;11:1102185. PMID: 37469694; https://doi.org/10.3389/fpubh.2023.1102185.
6. Vandevenne MM, Favuzza E, Veta M, et al. Artificial intelligence for detecting keratoconus. Cochrane Database Syst Rev. 2023 Nov 15;11(11):CD014911. PMID: 37965960; https://doi.org/10.1002/14651858.CD014911.pub2.

7. Kang C, Lo JE, Zhang H, et al. Inteligência artificial para diagnóstico de degeneração macular exsudativa relacionada à idade. Cochrane Database of Systematic Reviews. 2024 [cited 2024 Dec 19];10:CD015522. https://doi.org/10.1002/14651858.CD015522.pub2.
8. Chuchu N, Takwoingi Y, Dinnes J, et al. Smartphone applications for triaging adults with skin lesions that are suspicious for melanoma. Cochrane Database Syst Rev. 2018 Dec 4;12(12):CD013192. PMID: 30521685; https://doi.org/10.1002/14651858.CD013192.
9. Ferrante di Ruffano L, Takwoingi Y, Dinnes J, et al. Técnicas de diagnóstico assistidas por computador (baseadas em dermatoscopia e espectroscopia) para diagnóstico de câncer de pele em adultos. Cochrane Database of Systematic Reviews; 2018 [cited 2024 Dec 19];12:CD013186. https://doi.org/10.1002/14651858.CD013186.
10. Silva GG, Paixão HP, Rodrigues MLA. Desafios do uso da inteligência artificial nos diagnósticos de saúde: uma revisão integrativa. Cad Ibero Am Direito Sanit [Internet]. 2024 July 1 [cited 2025 Apr 14];13(2):11–8. Available from: https://www.cadernos.prodisa.fiocruz.br/index.php/cadernos/article/view/1241.
11. Lobo LC. Inteligência artificial e medicina. Rev Bras Educ Med. 2017 Apr-June;41(2). https://doi.org/10.1590/1981-52712015v41n2esp.
12. Lee EE, Torous J, De Choudhury M, et al. Artificial Intelligence for mental health care: clinical applications, barriers, facilitators, and artificial wisdom. Biol Psychiatry Cogn Neurosci Neuroimaging. 2021 Sep;6(9):856–64. Epub 2021 Feb 8. PMID: 33571718; https://doi.org/10.1016/j.bpsc.2021.02.001.
13. Amaro E Jr, Nakaya H, Rizzo LV. Inteligência artificial em saúde. Rev USP [Internet]. 2024 June 20 [cited 2025 Apr 14];(141):41–50. Disponível em: https://www.revistas.usp.br/revusp/article/view/225206