

EDITORA
MANUAL

INTELIGÊNCIA ARTIFICIAL NA SAÚDE

Diagnóstico, Decisão Clínica, Segurança do Paciente e Desafios Éticos

Ana Carla do Espírito Santo • Flávio Eduardo Fernandes Alves
José Peixoto de Alencar Neto • Márcia Silva Conceição

INTELIGÊNCIA ARTIFICIAL NA SAÚDE

Diagnóstico, Decisão Clínica, Segurança do Paciente e Desafios Éticos

Flávio Eduardo Fernandes Alves • Ana Carla do Espírito Santo
Márcia Silva Conceição • José Peixoto de Alencar Neto

Organização: Kátia Karolini Amaro El Alam

EDITORA MANUAL • 2026

INTELIGÊNCIA ARTIFICIAL NA SAÚDE

Diagnóstico, Decisão Clínica, Segurança do Paciente e Desafios Éticos

• • •

Flávio Eduardo Fernandes Alves • Ana Carla do Espírito Santo
Márcia Silva Conceição • José Peixoto de Alencar Neto

Organização: Kátia Karolini Amaro El Alam

EDITORA MANUAL • 1ª edição • 2026

EXPEDIENTE EDITORIAL

TÍTULO

Inteligência Artificial na Saúde: Diagnóstico, Decisão Clínica, Segurança do Paciente e Desafios Éticos

EDITORA

EDITORA MANUAL CNPJ: 12.072.139/0001-20

(27) 99750-4756 www.editoramanual.com.br editoramanualoficial@gmail.com

Rua Costa Brandão,785, Loja A, São Benedito, Cariacica, Espírito Santo – ES, CEP:

29145-340

EDIÇÃO

1ª edição, 2026

AUTORES

Flávio Eduardo Fernandes Alves

Ana Carla do Espírito Santo

Márcia Silva Conceição

José Peixoto de Alencar Neto

ORGANIZAÇÃO

Kátia Karolini Amaro El Alam

EDITOR-CHEFE

Afonso Henrique Souza de Assis

PRODUTOR EDITORIAL

Kátia Karolini Amaro El Alam

REVISÃO

Kátia Karolini Amaro El Alam

CAPA

Claude

BIBLIOTECÁRIA RESPONSÁVEL

Janaina Ramos

REVISÃO DE TEXTO E CAPA

Kátia Karolini Amaro El Alam

DIAGRAMAÇÃO
Kátia Karolini Amaro El Alam

CORPO EDITORIAL
Dra. Mônica Aparecida Bortolotti – Universidade Estadual do Paraná
Dr. Walmir Fernandes Pereira – MUST University
Dr. Luiz Gonzaga Lapa Junior – Universidade de Brasília
Dr. Gabriel Arruda Burani - Univerdidad de Ciencias Empresariales e Sociales
Dr. Januario Neto Pereira Sarmento – Instituto Federal de Educação, Ciência e Tecnologia do Tocantins
Dr. Tiago Piñeiro Martins – Secretaria Municipal de Educação de Salvador
Dra. Lucimara Cabreira Benitez – Universidade Estadual de Mato Grosso do Sul
Me. Afonso Henrique Souza de Assis – Universidad Leonardo da Vinci
Me. Elizabete Nascimento Aguiar – Universidade de Passo Fundo
Me. Swelen Freitas Gabarron – Universidade Tuiuti do Paraná
Me. José Neto de Oliveira Felippe – Faculdade de Caldas Novas
Me. Sueli Cristina Merotto Pereira – Secretaria Municipal de Educação - Prefeitura de Marataízes

Catalogação na publicação

Elaborada por Bibliotecária Janaina Ramos – CRB-8/9166

I61

Inteligência artificial na saúde: diagnóstico, decisão clínica, segurança do paciente e desafios éticos / Organização de Kátia Karolini Amaro El Alam. – Cariacica/ES: Editora Manual, 2026.

Autores: Ana Carla do Espírito Santo, Flávio Eduardo Fernandes Alves, José Peixoto de Alencar Neto, Márcia Silva Conceição.

Livro em PDF

ISBN: 978-65-831-2117-2
DOI: 10.29327/5932242

1. Inteligência artificial na saúde. 2. Diagnóstico. 3. Segurança do paciente. I. Alam, Kátia Karolini Amaro El (Organizadora). II. Título.

CDD 610.28

Índice para catálogo sistemático

I. Inteligência artificial na saúde

Obra publicada em Setembro de 2026.

DOI: 10.29327/5932242

© 2026. Todos os direitos reservados.

Nenhuma parte desta obra poderá ser reproduzida, armazenada ou transmitida por qualquer meio sem observância da legislação aplicável e dos direitos autorais.

APRESENTAÇÃO

Inteligência Artificial na Saúde: Diagnóstico, Decisão Clínica, Segurança do Paciente e Desafios Éticos examina uma transformação que já ultrapassou o campo das promessas tecnológicas. Sistemas capazes de reconhecer padrões, estimar riscos, apoiar decisões, interpretar imagens, organizar informações clínicas e produzir linguagem passaram a integrar diferentes etapas do cuidado. A questão central da obra não é, portanto, se a inteligência artificial pode participar da assistência em saúde, mas sob quais condições essa participação produz benefício clínico real sem ampliar riscos, desigualdades ou zonas de responsabilidade indefinida.

O percurso parte da história da inteligência artificial em medicina e acompanha a passagem de sistemas baseados em regras para métodos de aprendizagem de máquina, modelos profundos e arquiteturas generativas e multimodais. Diagnóstico, imagem, prognóstico, medicina personalizada, monitoramento remoto e saúde populacional são analisados não apenas pelo desempenho técnico das ferramentas, mas pela forma como resultados computacionais entram em fluxos assistenciais concretos, interagem com profissionais e influenciam decisões cujas consequências recaem sobre pessoas.

A obra dedica atenção especial à distância entre acurácia experimental e segurança clínica. Validação externa, mudança de distribuição dos dados, monitoramento pós-implementação, vieses, desempenho desigual entre grupos e risco de automação excessiva mostram por que bons resultados em bancos de teste não bastam para justificar adoção. A segurança do paciente exige compreender onde o sistema funciona, onde falha, quem reconhece o erro, como a decisão pode ser contestada e quais mecanismos institucionais permitem interromper ou corrigir um uso inadequado.

Privacidade, explicabilidade, cibersegurança, regulação e responsabilidade profissional compõem outra dimensão indispensável. Dados de saúde são simultaneamente matéria-prima para inovação e informações sensíveis cuja exposição pode gerar danos duradouros. Por isso, governança não aparece como etapa burocrática posterior à implantação, mas como parte do próprio desenho do cuidado: inventariar sistemas, definir finalidades, delimitar autonomia, registrar versões, acompanhar incidentes, distribuir responsabilidades e estabelecer critérios de revisão são condições para que a tecnologia permaneça subordinada às necessidades clínicas e aos direitos dos pacientes.

Destinada a profissionais e estudantes das áreas da saúde, pesquisadores, gestores e equipes envolvidas com transformação digital, a obra sustenta uma perspectiva de cuidado aumentado, e não simplesmente automatizado. O futuro mais confiável não depende de sistemas que pareçam mais humanos, mas de tecnologias que ampliem capacidades sem dissolver julgamento profissional, vínculo, equidade, transparência e prestação de contas. A inteligência artificial torna-se clinicamente valiosa quando sua utilidade pode ser demonstrada, seus limites permanecem visíveis e instituições e profissionais continuam capazes de responder pelas decisões tomadas com seu apoio.

SUMÁRIO

INTRODUÇÃO 9

CAPÍTULO 1 DA LÓGICA SIMBÓLICA AOS MODELOS GENERATIVOS: COMO A INTELIGÊNCIA ARTIFICIAL CHEGOU À SAÚDE 16

CAPÍTULO 2 QUANDO O ALGORITMO APRENDE A RECONHECER DOENÇA: DIAGNÓSTICO, IMAGEM E DETECÇÃO DE PADRÕES 26

CAPÍTULO 3 DA PREVISÃO À DECISÃO: SUPORTE CLÍNICO, PROGNÓSTICO E MEDICINA PERSONALIZADA 36

CAPÍTULO 4 CUIDADO CONTÍNUO E SISTEMAS INTELIGENTES: MONITORAMENTO, DISPOSITIVOS, ATENÇÃO REMOTA E SAÚDE POPULACIONAL 47

CAPÍTULO 5 QUANDO A IA PASSA A PRODUZIR: MODELOS GENERATIVOS, MULTIMODAIS E LINGUAGEM CLÍNICA 57

CAPÍTULO 6 QUANDO DESEMPENHO NÃO BASTA: VALIDAÇÃO CLÍNICA, SEGURANÇA DO PACIENTE E MONITORAMENTO PÓS-IMPLEMENTAÇÃO 67

CAPÍTULO 7 QUEM FICA DE FORA DOS DADOS? VIÉS ALGORÍTMICO, EQUIDADE E DESIGUALDADES EM SAÚDE 77

CAPÍTULO 8 DADOS QUE CUIDAM E DADOS QUE EXPÕEM: PRIVACIDADE, TRANSPARÊNCIA, EXPLICABILIDADE E CIBERSEGURANÇA 86

CAPÍTULO 9 QUEM RESPONDE PELA DECISÃO? REGULAÇÃO, RESPONSABILIDADE PROFISSIONAL E GOVERNANÇA DA IA EM SAÚDE 96

CAPÍTULO 10 CUIDADO AUMENTADO, NÃO SUBSTITUÍDO: COLABORAÇÃO HUMANO-IA E CAMINHOS PARA UMA SAÚDE CONFIÁVEL E SUSTENTÁVEL 106

CONSIDERAÇÕES FINAIS 115

REFERÊNCIAS 117

SOBRE OS AUTORES 126

INTRODUÇÃO

Durante boa parte da história da medicina, a inteligência foi tratada como atributo quase inseparável da experiência humana. Reconhecer uma doença a partir de sinais incompletos, ponderar hipóteses concorrentes, interpretar uma imagem, escolher entre tratamentos possíveis e explicar uma decisão a quem está vulnerável pareciam tarefas cuja complexidade residia justamente na combinação de conhecimento, julgamento e responsabilidade. A entrada da inteligência artificial nesse território não eliminou essa complexidade. Ao contrário, tornou mais visível uma pergunta que durante algum tempo foi obscurecida pelo entusiasmo tecnológico: o que significa, de fato, dizer que uma máquina pode participar de uma decisão em saúde?

Essa pergunta é mais difícil do que comparar a acurácia de um algoritmo com a de um profissional. Um sistema pode classificar imagens com desempenho elevado em uma base de teste e ainda falhar quando encontra pacientes diferentes daqueles usados no desenvolvimento. Pode produzir uma resposta linguisticamente convincente e, ao mesmo tempo, omitir uma contraindicação importante. Pode reduzir minutos de uma tarefa administrativa e aumentar horas de verificação humana. Pode identificar um padrão que escaparia ao olhar clínico e, em outra situação, reproduzir desigualdades existentes nos próprios dados. A inteligência artificial em saúde não é apenas um problema de capacidade computacional. É uma intervenção sociotécnica: seu efeito depende do modelo, dos dados, do profissional, do paciente, do fluxo de trabalho, da organização e das regras que definem como a tecnologia será utilizada.

Essa mudança de perspectiva é necessária porque o debate público costuma oscilar entre duas imagens igualmente pobres. Em uma delas, a inteligência artificial aparece como solução inevitável para sistemas de saúde pressionados por custos, escassez de profissionais, excesso de informação e demanda crescente. Na outra, surge como ameaça destinada a substituir médicos, desumanizar o cuidado e produzir

decisões opacas. As duas imagens têm força retórica, mas pouca capacidade explicativa. Tecnologias não entram na assistência como personagens autônomos. Elas são projetadas para determinadas finalidades, treinadas com determinados conjuntos de dados, incorporadas por organizações específicas e usadas por pessoas que continuam responsáveis por interpretar resultados e responder às consequências.

A história da inteligência artificial em medicina ajuda a perceber por que essa cautela não é sinal de resistência à inovação. Desde os primeiros sistemas especialistas, pesquisadores tentam formalizar partes do raciocínio clínico em estruturas computacionais. MYCIN, desenvolvido nos anos 1970 para apoiar decisões relacionadas a infecções bacterianas e antimicrobianos, tornou-se um marco porque demonstrou que regras codificadas podiam produzir recomendações sofisticadas em um domínio restrito. INTERNIST-I, apresentado no início dos anos 1980 como consultor computadorizado para medicina interna, seguiu outra estratégia, trabalhando com relações entre achados e diagnósticos em larga escala. Esses sistemas não se converteram em rotina clínica ampla, mas mostraram algo que continua atual: bom desempenho técnico não garante adoção. Integração ao trabalho, atualização do conhecimento, explicabilidade, responsabilidade e confiança sempre estiveram no centro do problema (MILLER; POPLE; MYERS, 1982; KAUL; ENSLIN; GROSS, 2020).

Nas décadas seguintes, a trajetória da inteligência artificial foi marcada por alternância entre expectativas elevadas e períodos de frustração. Limitações de poder computacional, disponibilidade de dados, dificuldade de representar conhecimento e custo de manter regras atualizadas restringiram os sistemas baseados principalmente em lógica simbólica. A transformação veio gradualmente com a digitalização de registros, exames e imagens; com o aumento da capacidade de armazenamento e processamento; e com o amadurecimento de métodos estatísticos capazes de aprender regularidades diretamente dos dados. O foco deslocou-se. Em vez de perguntar apenas como programar explicitamente o conhecimento de um especialista, tornou-se

possível perguntar que padrões um modelo conseguiria aprender a partir de milhares ou milhões de exemplos.

A ascensão do aprendizado de máquina e, mais tarde, do deep learning alterou a escala dessas possibilidades. Imagens médicas ofereceram um terreno particularmente fértil porque radiografias, tomografias, ressonâncias, lâminas histopatológicas e fotografias clínicas já eram produzidas em formato digital e podiam ser associadas a rótulos diagnósticos. Redes neurais profundas passaram a demonstrar desempenho relevante em tarefas de classificação e detecção, enquanto técnicas semelhantes se expandiram para registros eletrônicos, genômica, sinais fisiológicos e previsão de risco. Esse movimento fez a inteligência artificial sair progressivamente do laboratório de informática médica e entrar no cotidiano regulatório, empresarial e assistencial (YU; BEAM; KOHANE, 2018; ESTEVA et al., 2019; TOPOL, 2019; RAJPURKAR et al., 2022).

Ainda assim, uma das lições mais importantes dessa fase foi a distância entre prova de conceito e benefício clínico. Modelos podem funcionar muito bem em dados retrospectivos cuidadosamente selecionados e perder desempenho quando mudam prevalência, equipamento, protocolo, população ou ambiente de atendimento. Acurácia média pode esconder erros concentrados em determinados grupos. Métricas estatísticas podem não refletir o custo clínico de falsos positivos e falsos negativos. E uma ferramenta tecnicamente superior pode não produzir benefício se chegar ao profissional no momento errado ou exigir uma forma de interação incompatível com o fluxo real. A avaliação da inteligência artificial em saúde precisa, portanto, ultrapassar a pergunta “o algoritmo acerta?” e alcançar questões sobre utilidade, segurança, generalização e impacto na prática.

A partir de 2022 e 2023, outra mudança ganhou enorme visibilidade: modelos de linguagem de grande escala e modelos fundacionais passaram a produzir texto, imagens e outras saídas com flexibilidade que não dependia de uma única tarefa previamente delimitada. Na medicina, esse novo paradigma sugeriu sistemas capazes

de interpretar combinações de prontuários, imagens, exames laboratoriais e linguagem natural, além de resumir documentos, responder perguntas e apoiar atividades clínicas. Moor et al. (2023) descreveram esse horizonte como inteligência artificial médica generalista, mas também ressaltaram que modelos desse tipo desafiam formas tradicionais de validação e regulação. A mesma flexibilidade que amplia possibilidades dificulta definir antecipadamente todos os usos e todos os modos de falha.

A Organização Mundial da Saúde reagiu a essa transformação em duas etapas importantes. Em 2021, publicou orientação global sobre ética e governança da inteligência artificial em saúde, propondo princípios como proteção da autonomia humana, promoção de segurança e bem-estar, transparência, responsabilidade, inclusão, equidade e sustentabilidade. Em 2024, diante da expansão dos grandes modelos multimodais, publicou orientação específica para esse tipo de tecnologia, chamando atenção para riscos ligados a informações falsas, vieses, automação inadequada, uso de dados, cibersegurança e necessidade de governança ao longo de todo o ciclo de vida (OMS, 2021; OMS, 2024). Essas recomendações ajudam a deslocar o debate de uma pergunta sobre poder computacional para uma pergunta sobre condições legítimas de uso.

Esse deslocamento já aparece também na regulação. A Food and Drug Administration mantém uma lista crescente de dispositivos médicos com componentes de inteligência artificial autorizados nos Estados Unidos e vem consolidando uma abordagem de ciclo de vida para esses produtos, incluindo planos de mudança predeterminados e monitoramento de desempenho. Na União Europeia, o Regulamento de Inteligência Artificial estabelece um regime baseado em risco e inclui, em determinadas condições, sistemas incorporados a produtos regulados como dispositivos médicos entre os sistemas de alto risco. No Brasil, o tema deixou de ocupar apenas o campo da inovação e passou a receber disciplina profissional específica: a Resolução CFM nº 2.454/2026 normatiza pesquisa, desenvolvimento, governança, auditoria, monitoramento, capacitação e uso responsável de soluções de

inteligência artificial na medicina. Esses marcos diferem em alcance e natureza jurídica, mas convergem em um ponto: a tecnologia precisa ser avaliada não apenas antes de entrar no cuidado, mas durante seu uso e suas mudanças. (UNIÃO EUROPEIA, 2024; CFM, 2026).

A existência de regulação não elimina a incerteza. Modelos adaptativos podem mudar, dados podem envelhecer, populações podem se transformar e novos usos podem surgir depois da implantação. Além disso, sistemas generativos introduzem um tipo de erro particularmente difícil: uma resposta pode ser coerente, bem escrita e falsa. A revisão sistemática de Busch et al. (2025), ao mapear aplicações de grandes modelos de linguagem no cuidado ao paciente, identificou usos em resposta a perguntas, produção de informação, sumarização, tradução e documentação, mas também limitações relacionadas a incorreção, incompletude, vieses, falta de transparência e segurança. Estudos experimentais em decisão clínica realista reforçam que desempenho em provas ou questões fechadas não basta para demonstrar segurança em ambientes nos quais informações chegam de forma incompleta, mudam ao longo do tempo e exigem aderência a protocolos (HAGER et al., 2024).

É justamente nesse intervalo entre potencial e confiabilidade que este livro se posiciona. A obra não parte da premissa de que inteligência artificial é intrinsecamente boa ou ruim para a saúde. Tampouco pretende tratar profissionais como obstáculos à inovação ou pacientes como receptores passivos de decisões automatizadas. O objetivo é compreender o que muda quando sistemas computacionais passam a participar de atividades historicamente relacionadas ao julgamento clínico, à vigilância, à comunicação e à organização do cuidado.

A tese que orienta os capítulos é simples de formular, mas exigente de aplicar: uma inteligência artificial só deve ser considerada clinicamente valiosa quando seu benefício permanece defensável depois que saímos do laboratório e incluímos o paciente real, o profissional real, a instituição real e as consequências reais do erro. Isso significa que desempenho algorítmico é condição importante, mas insuficiente.

Precisamos examinar como dados foram produzidos, para quem o modelo funciona, quem identifica sua falha, como a decisão pode ser contestada, quais informações chegam ao usuário, como o sistema é monitorado e quem responde quando algo dá errado.

Por essa razão, a sequência dos capítulos acompanha uma progressão deliberada. Primeiro reconstruímos a trajetória que levou a inteligência artificial da lógica simbólica aos modelos generativos. Em seguida, entramos nas aplicações que mais contribuíram para sua credibilidade clínica: diagnóstico, análise de imagens e reconhecimento de padrões. Depois, o foco passa da identificação de doença para previsão e decisão terapêutica, e daí para monitoramento contínuo, dispositivos e saúde populacional. Somente depois desse percurso examinamos em profundidade modelos generativos e multimodais, porque suas possibilidades e riscos se tornam mais compreensíveis quando comparados às formas anteriores de inteligência artificial.

A segunda metade da obra muda de direção. Em vez de perguntar o que a tecnologia consegue fazer, perguntamos sob quais condições pode ser confiável. Validação clínica, segurança do paciente, mudança de distribuição dos dados, automação indevida, vieses, equidade, privacidade, explicabilidade, cibersegurança e responsabilidade regulatória entram como partes de um mesmo problema. O livro termina retornando àquilo que tecnologia alguma elimina: o cuidado é uma relação em que decisões têm consequências humanas e precisam permanecer vinculadas a responsabilidade, possibilidade de contestação e compromisso com o paciente.

A inteligência artificial pode ampliar a capacidade de enxergar padrões, organizar informação e antecipar riscos. Pode também criar novas formas de erro, dependência e desigualdade. Entre essas duas possibilidades existe um trabalho institucional e científico que não pode ser pulado. O futuro da inteligência artificial na saúde será definido menos pela pergunta sobre se máquinas conseguirão realizar tarefas médicas e mais pela capacidade de construir sistemas em que tecnologia aumente o cuidado sem diminuir a responsabilidade.

CAPÍTULO 1
DA LÓGICA SIMBÓLICA AOS MODELOS GENERATIVOS: COMO A INTELIGÊNCIA ARTIFICIAL CHEGOU À SAÚDE

Quando raciocinar significava escrever regras

A história da inteligência artificial em saúde costuma ser narrada como uma sucessão de avanços tecnológicos: primeiro os sistemas especialistas, depois o aprendizado de máquina, em seguida as redes neurais profundas e, mais recentemente, os modelos generativos. Essa linha do tempo é útil, mas incompleta. O que realmente mudou ao longo das décadas foi a forma de imaginar onde o conhecimento clínico estaria localizado. Nos primeiros sistemas, ele precisava ser explicitado em regras; nos modelos estatísticos, passou a ser inferido a partir de dados; nos modelos fundacionais, parte desse conhecimento aparece distribuída em representações aprendidas em escala tão grande que nem sempre é possível reconstruir de maneira simples como uma resposta foi produzida. A evolução técnica, portanto, foi também uma mudança na relação entre conhecimento, explicação e controle.

A pergunta sobre máquinas capazes de realizar atividades associadas à inteligência antecede a medicina computacional. Em 1950, Alan Turing propôs deslocar a discussão abstrata sobre se máquinas "pensam" para um problema observável de comportamento. Poucos anos depois, a conferência de Dartmouth consolidou inteligência artificial como campo de investigação. Na saúde, entretanto, o desafio assumiu desde cedo uma particularidade: não bastava resolver problemas formais. Era necessário representar conhecimento clínico, lidar com incerteza e produzir recomendações em contextos nos quais um erro poderia afetar diretamente uma pessoa.

Os sistemas especialistas das décadas de 1960 e 1970 nasceram dessa tentativa. Em vez de aprenderem sozinhos a partir de grandes bases, eles dependiam de engenheiros de conhecimento e especialistas que transformavam raciocínio clínico em regras explícitas. A vantagem era intuitiva: se o sistema recomendava determinado caminho, seria possível inspecionar quais regras foram acionadas. A dificuldade aparecia na escala. A medicina contém exceções, ambiguidades, relações probabilísticas e conhecimento em constante mudança. Quanto mais complexo o domínio, maior o número de regras necessário para representá-lo.

MYCIN tornou-se um símbolo dessa fase. Desenvolvido em Stanford durante os anos 1970, o sistema utilizava regras para apoiar a identificação de infecções bacterianas e a escolha de antimicrobianos. Seu valor histórico não está em ter se tornado produto assistencial de uso generalizado - isso não ocorreu - mas em demonstrar que um programa poderia combinar informações clínicas, trabalhar com incerteza e fornecer justificativas para suas recomendações. Ele aproximou a computação de um problema que até então parecia profundamente dependente de expertise humana (KAUL; ENSLIN; GROSS, 2020).

INTERNIST-I ampliou a ambição. Publicado em 1982, o sistema foi descrito como consultor computadorizado para medicina interna geral. Seu conhecimento relacionava centenas de doenças a milhares de manifestações, tentando organizar diagnósticos diferenciais a partir de achados clínicos. O projeto mostrou que era possível construir bases extensas de conhecimento médico, mas também revelou o custo de mantê-las atualizadas e a dificuldade de reproduzir a flexibilidade com que um clínico atribui pesos diferentes ao mesmo achado dependendo do contexto (MILLER; POPLE; MYERS, 1982).

Esses projetos são importantes porque desmontam a ideia de que as dificuldades atuais da inteligência artificial surgiram apenas com modelos modernos. Explicabilidade, atualização, integração ao fluxo e responsabilidade já estavam presentes, ainda que sob outra forma. Um sistema baseado em regras era mais transparente em sua arquitetura, mas poderia falhar porque o conhecimento codificado estava incompleto. Um modelo atual pode aprender padrões que nenhum especialista escreveu, porém tornar mais difícil compreender por que uma saída específica foi produzida. Em ambos os casos, a tecnologia carrega escolhas humanas sobre quais informações importam e como serão utilizadas.

A fase dos sistemas especialistas também ensinou uma lição sobre a diferença entre demonstrar capacidade e construir utilidade. Um programa pode produzir uma recomendação tecnicamente interessante e não encontrar lugar na assistência. Para ser usado, precisa receber dados no momento certo, conversar com outros sistemas, respeitar rotinas, ser atualizado e contar com profissionais dispostos a confiar em seus resultados. A medicina não é apenas um conjunto de diagnósticos; é um processo social de coleta de informação, comunicação, decisão e responsabilização.

Quando essas limitações se somaram à capacidade computacional ainda restrita e à pequena disponibilidade de dados digitais, parte do entusiasmo diminuiu. A história

geral da inteligência artificial passou por períodos que mais tarde seriam descritos como “invernos da IA”, marcados por redução de financiamento e expectativas. Na medicina, entretanto, a pesquisa não desapareceu. Ela mudou de direção. Métodos probabilísticos, sistemas de apoio à decisão e aprendizado de máquina começaram a ocupar um espaço maior à medida que hospitais e laboratórios produziam volumes crescentes de dados eletrônicos.

Essa transição é decisiva porque muda a fonte do conhecimento do sistema. Na lógica simbólica, especialistas precisavam antecipar relações e transformá-las em regras. No aprendizado de máquina, o objetivo passa a ser estimar uma função a partir de exemplos. Se milhares de pacientes possuem determinados atributos e resultados conhecidos, um modelo pode aprender associações úteis para classificar novos casos. O conhecimento deixa de ser inteiramente escrito e passa a ser parcialmente extraído dos dados. Essa mudança amplia a escala, mas introduz uma dependência que atravessará todo o restante do livro: o que o modelo aprende depende do mundo que os dados conseguiram registrar.

Essa dependência é frequentemente resumida por uma expressão simples - “garbage in, garbage out” -, mas na saúde o problema é mais sofisticado. Dados podem ser tecnicamente corretos e ainda refletir desigualdades de acesso, diferenças de protocolo, decisões históricas ou práticas institucionais. Uma base de internações representa quem chegou ao hospital, não necessariamente todos os que adoeceram. Um banco de imagens pode concentrar equipamentos de determinados fabricantes. Um prontuário registra aquilo que profissionais tiveram tempo, motivo e recursos para documentar. O dado não é uma fotografia neutra da realidade; é produto de um processo de cuidado.

Quando o aprendizado de máquina passou a ocupar o centro da inteligência artificial médica, portanto, o problema de representação não desapareceu. Apenas mudou de forma. Antes, perguntava-se se as regras escritas representavam adequadamente o conhecimento clínico. Agora, tornou-se necessário perguntar se os dados representam adequadamente os pacientes e as situações em que o sistema será usado. Essa troca de pergunta abriria caminho para a fase seguinte, em que quantidade de dados e capacidade computacional cresceriam a ponto de permitir que modelos descobrissem representações cada vez mais complexas sem que cada característica precisasse ser definida antecipadamente.

Quando os dados passaram a ensinar a máquina

O avanço do deep learning não ocorreu de maneira isolada. Ele dependeu de uma infraestrutura que vinha sendo construída havia anos: digitalização de imagens, registros eletrônicos, bancos de dados clínicos, armazenamento em nuvem, unidades de processamento gráfico e redes capazes de compartilhar grandes volumes de informação. Quando esses elementos se encontraram com novas arquiteturas de redes neurais, tarefas que exigiam intensa engenharia manual de características passaram a ser abordadas de outra forma. O modelo poderia aprender representações diretamente de pixels, sinais ou sequências.

Na medicina, a imagem foi um dos primeiros espaços em que essa mudança se tornou visível. Radiologia, oftalmologia, dermatologia e patologia reúnem tarefas nas quais profissionais reconhecem padrões visuais associados a doenças. Isso não significa que diagnóstico seja apenas reconhecimento de imagem, mas cria um problema computacional bem definido: dado um conjunto de exames e rótulos, é possível treinar um modelo para identificar padrões associados a determinadas condições. A disponibilidade de grandes bases e o amadurecimento de redes convolucionais tornaram esse problema tecnicamente tratável em escala antes impensável (ESTEVA et al., 2019).

O interesse cresceu porque alguns estudos passaram a demonstrar desempenhos comparáveis aos de especialistas em tarefas restritas. O impacto simbólico dessas comparações foi enorme. Manchetes sugeriam que algoritmos “diagnosticavam como médicos” ou “superavam especialistas”, mesmo quando o estudo avaliava apenas uma classificação delimitada. Topol (2019) chamou atenção para uma possibilidade mais produtiva: a convergência entre inteligência humana e artificial, na qual sistemas computacionais assumiriam tarefas de reconhecimento e processamento enquanto profissionais preservariam dimensões relacionais, contextuais e decisórias que não se reduzem à classificação.

Essa distinção continua essencial. Identificar uma lesão em uma imagem é parte do diagnóstico, não o diagnóstico inteiro. O profissional precisa relacionar o achado à história, aos sintomas, aos exames prévios, à probabilidade pré-teste e às consequências de agir ou não agir. Mesmo em áreas altamente visuais, a imagem existe dentro de um percurso clínico. Um algoritmo pode aumentar sensibilidade para determinada alteração e, ainda assim, gerar excesso de investigações se a prevalência for baixa ou se sua saída for interpretada sem contexto.

O mesmo raciocínio vale para previsão de risco. Modelos podem estimar probabilidade de deterioração, reinternação, mortalidade ou resposta a tratamento usando grandes conjuntos de variáveis. A capacidade de combinar milhares de sinais é atraente porque nenhum ser humano consegue processar simultaneamente a mesma quantidade de dados. Mas previsão não é decisão. Saber que um paciente pertence a um grupo de risco só produz benefício quando existe uma ação adequada, disponível e realizada no momento oportuno. Um alerta que ninguém consegue atender pode aumentar fadiga e reduzir atenção aos casos realmente importantes.

Essa diferença entre performance e impacto passou a ocupar espaço crescente na literatura. Yu, Beam e Kohane (2018) já alertavam que a expansão da inteligência artificial biomédica trazia desafios econômicos, sociais, legais e de integração. Rajpurkar et al. (2022), ao revisar avanços em inteligência artificial para saúde e medicina, destacaram a passagem gradual de estudos retrospectivos para avaliações prospectivas e a importância da colaboração humano-IA. O campo amadurecia justamente ao perceber que comparar curvas ROC em bases estáticas era apenas uma etapa inicial.

Uma ferramenta de inteligência artificial entra em um sistema de saúde que já possui pessoas, prioridades, falhas e desigualdades. Se o algoritmo detecta mais casos, o serviço precisa ter capacidade para absorver exames adicionais. Se prioriza imagens suspeitas, a equipe precisa entender como a fila foi reorganizada. Se gera um escore de risco, alguém precisa saber qual conduta corresponde a cada faixa. A tecnologia não apenas informa o trabalho; pode reorganizá-lo. Por isso, avaliar seu impacto exige observar o que acontece depois da saída do modelo.

Outra mudança importante dessa fase foi regulatória. A inteligência artificial deixou de ser apenas objeto de publicação acadêmica e tornou-se componente de dispositivos colocados no mercado. A FDA mantém uma lista pública de dispositivos médicos habilitados por inteligência artificial e continua atualizando-a, com forte presença de aplicações de radiologia. A própria necessidade de uma lista específica mostra que o campo já ultrapassou a etapa em que IA médica era essencialmente experimental. A questão regulatória passou a incluir como avaliar atualizações, mudanças de desempenho, transparência para usuários e monitoramento ao longo da vida útil do produto (FDA, 2026a).

Esse desafio é diferente daquele enfrentado por um software tradicional. Um programa determinístico tende a executar a mesma lógica quando recebe a mesma

entrada. Sistemas baseados em aprendizado podem depender de distribuições de dados e, em alguns cenários, ser atualizados para melhorar desempenho. Isso criou a necessidade de pensar ciclo de vida. Em 2025, a FDA consolidou orientação sobre planos predeterminados de controle de mudanças para funções de software de dispositivos habilitados por IA, permitindo que determinadas modificações sejam antecipadas e avaliadas dentro de limites definidos sem tratar cada evolução como se fosse um produto completamente novo. O movimento regulatório reconhece a natureza iterativa da tecnologia, mas também deixa claro que "aprender" não pode significar mudar sem controle.

A expansão do deep learning também trouxe de volta um problema que sistemas especialistas pareciam resolver melhor: explicação. Redes profundas podem produzir previsões altamente precisas sem fornecer uma cadeia de raciocínio equivalente às regras de um sistema simbólico. Métodos de interpretabilidade tentam mostrar regiões de imagem, variáveis ou características mais associadas a uma saída, mas essas ferramentas não são sinônimo de explicação clínica completa. Uma visualização pode indicar onde o modelo "olhou" sem demonstrar por que aquela evidência deveria ser considerada causal ou clinicamente relevante.

Essa tensão entre performance e inteligibilidade ajudou a deslocar a ideia de confiança. Confiar não precisa significar entender cada operação matemática interna, assim como profissionais utilizam exames laboratoriais sem conhecer cada detalhe do equipamento. Mas a opacidade aumenta a necessidade de conhecer outros elementos: população de validação, finalidade pretendida, limites, taxa de erro, condições de uso e mecanismos de monitoramento. Em sistemas de alto impacto, confiança não pode ser construída apenas pela reputação do desenvolvedor ou pela fluência da interface.

A fase do deep learning, portanto, aproximou a inteligência artificial da prática ao mesmo tempo que tornou mais evidentes os limites do olhar puramente técnico. Quanto mais o modelo se aproximava da assistência, mais importavam questões que não estavam dentro do código. O próximo salto - dos modelos especializados para os modelos fundacionais e generativos - ampliaria essa tensão. Em vez de uma ferramenta treinada para uma tarefa, surgiriam sistemas capazes de realizar muitas tarefas, conversar em linguagem natural e produzir respostas que parecem compreensão. A fronteira entre ferramenta e interlocutor ficaria mais difícil de perceber.

Quando a máquina passou a conversar

Os modelos de linguagem de grande escala transformaram a percepção pública da inteligência artificial porque mudaram a interface. Durante décadas, muitos sistemas médicos operavam nos bastidores: classificavam uma imagem, geravam um risco ou apoiavam um software específico. Com os modelos generativos, o usuário passou a formular perguntas em linguagem natural e receber respostas igualmente naturais. A tecnologia deixou de parecer apenas instrumento de análise e começou a ocupar um espaço discursivo. Na saúde, onde explicações, perguntas e decisões são construídas por linguagem, essa mudança possui consequências profundas.

A arquitetura dos transformers, o treinamento em grandes corpora e a capacidade de realizar múltiplas tarefas a partir de instruções permitiram o surgimento de modelos fundacionais reutilizáveis. Moor et al. (2023) propuseram a ideia de uma inteligência artificial médica generalista capaz de combinar modalidades como texto clínico, imagens, exames laboratoriais e genômica e produzir saídas diversas. O conceito é uma proposta de paradigma, não prova de que sistemas atuais possuam competência geral. Sua importância está em antecipar uma mudança: o mesmo modelo pode ser adaptado a tarefas que antes exigiriam ferramentas separadas.

Essa flexibilidade abre possibilidades evidentes. Um modelo pode resumir prontuários extensos, transformar linguagem técnica em explicações mais acessíveis, auxiliar na elaboração de documentação, sugerir hipóteses, organizar resultados de exames, apoiar pesquisa e facilitar comunicação entre diferentes formatos de informação. Em contextos de alta carga administrativa, a promessa de reduzir tempo documental é especialmente atraente. Em ambientes com escassez de especialistas, sistemas capazes de organizar conhecimento podem também ampliar acesso a informação.

Mas a linguagem natural introduz uma forma particular de risco: credibilidade estilística. Uma classificação tradicional costuma apresentar um escore ou rótulo. Um modelo generativo constrói frases completas, oferece justificativas e pode adaptar o tom ao interlocutor. Essa fluência torna a saída mais fácil de usar e, ao mesmo tempo, mais difícil de desconfiar. Uma informação incorreta pode parecer tão segura quanto uma informação correta. O problema não é apenas a existência de “alucinações”, mas a ausência de um sinal confiável que permita ao usuário perceber quando o modelo ultrapassou aquilo que seus dados ou seu treinamento sustentam.

A orientação da OMS para grandes modelos multimodais enfatiza exatamente essa tensão. O documento reconhece usos potenciais na assistência, pesquisa, educação, saúde pública e desenvolvimento de medicamentos, mas propõe governança voltada a segurança, transparência, equidade, responsabilidade e avaliação independente. A preocupação é ampliada porque modelos generalistas podem ser usados fora de finalidades cuidadosamente validadas. Um sistema treinado para responder perguntas gerais pode ser convidado por usuários a interpretar sintomas, sugerir tratamentos ou avaliar urgência sem que tenha sido desenvolvido e testado para essas decisões (OMS, 2024).

A revisão sistemática de Busch et al. (2025) mostra como essa expansão já aparece na literatura. Entre estudos de modelos de linguagem no cuidado ao paciente, predominavam aplicações para responder perguntas médicas, produzir ou adaptar informações, resumir e traduzir textos e apoiar documentação. As limitações incluíam falta de otimização para domínios específicos, baixa transparência, não reprodutibilidade, respostas incompletas, incorretas, inseguras ou enviesadas. O dado importante não é apenas que modelos erram. É que diferentes tipos de erro exigem diferentes estratégias de mitigação.

Uma resposta fabricada pode exigir verificação factual; uma omissão pode passar despercebida porque nada parece errado; uma recomendação desatualizada pode refletir dados antigos; um viés pode aparecer apenas em determinados grupos; uma resposta excessivamente confiante pode aumentar automação indevida. Colocar todos esses problemas sob uma única categoria empobrece a análise. Segurança clínica exige conhecer o modo de falha, sua frequência e sua consequência.

Hager et al. (2024) demonstraram essa diferença ao avaliar grandes modelos de linguagem em tarefas de decisão clínica mais realistas. Os autores observaram que desempenho em exames e benchmarks não captava adequadamente habilidades necessárias à prática, como solicitar informação relevante, seguir diretrizes e manter robustez diante de mudanças na formulação da instrução. Modelos podiam produzir diagnósticos e recomendações de tratamento inadequados quando recebiam informações incompletas e exigiam supervisão contínua. A mensagem não é que grandes modelos sejam inúteis, mas que a forma de avaliá-los precisa se aproximar do ambiente em que se pretende usá-los.

Essa questão muda o sentido da palavra “assistente”. Um assistente clínico humano não é avaliado apenas por responder perguntas de prova; aprende contexto,

reconhece limites, pede ajuda, acompanha consequências e está inserido em uma cadeia de responsabilidade. Chamar um modelo de "assistente" pode criar expectativas que o sistema não consegue cumprir. A linguagem de produto influencia a forma como profissionais e pacientes interpretam autoridade.

Por isso, supervisão humana precisa ser entendida com mais profundidade do que a presença formal de um profissional no circuito. Se o sistema produz centenas de saídas e o humano apenas confirma rapidamente cada uma, a supervisão pode existir no organograma e falhar na prática. Automação pode deslocar atenção, criar viés de confirmação e reduzir disposição para revisar recomendações. O desenho da interface, o tempo disponível e a clareza sobre incerteza determinam se o profissional realmente mantém controle.

A Resolução CFM nº 2.454/2026 insere essa preocupação no contexto brasileiro ao estabelecer princípios para uso responsável de IA na medicina e tratar temas como governança, auditoria, monitoramento, capacitação e responsabilidade. A existência de uma norma profissional específica mostra que a inteligência artificial deixou de ser um tema distante do exercício da medicina. A partir de 2026, a discussão no Brasil precisa considerar não apenas o que sistemas conseguem fazer, mas como seu uso se relaciona ao dever profissional, à autonomia e à segurança do paciente (CFM, 2026).

A regulação sanitária brasileira também oferece um ponto importante. A Anvisa enquadra software conforme sua finalidade médica, e a RDC nº 657/2022 disciplina software como dispositivo médico. Em orientações técnicas, a Agência já indicava que o uso de inteligência artificial ou aprendizado de máquina não define sozinho se um produto é um dispositivo; o enquadramento depende da finalidade pretendida. Essa distinção evita regular "IA" como categoria abstrata e concentra atenção no que o sistema faz no cuidado. Em 2026, a própria agenda regulatória da Anvisa inclui revisão das regras de software como dispositivo médico, sinal de que o marco continua evoluindo diante de tecnologias que se transformam rapidamente. (ANVISA, 2022b; ANVISA, 2026).

A entrada de modelos generativos, portanto, não inaugura uma ruptura total com o passado. Ela intensifica problemas antigos. Sistemas especialistas já exigiam conhecimento correto; modelos estatísticos já dependiam da população de treinamento; deep learning já levantava dúvidas sobre explicabilidade; software médico já exigia validação e controle. O que os modelos generativos acrescentam é

amplitude, velocidade, acessibilidade e linguagem. Isso aumenta tanto a utilidade potencial quanto a superfície de risco.

A história percorrida neste capítulo revela um padrão. Cada geração de inteligência artificial resolveu parte das limitações da anterior e criou novos tipos de incerteza. Regras explícitas eram limitadas pela dificuldade de codificar conhecimento; aprendizado de máquina ampliou escala, mas tornou os dados centrais; deep learning melhorou representação, mas aprofundou opacidade; modelos fundacionais ampliaram flexibilidade, mas dificultaram delimitar uso e prever todos os modos de falha. O progresso não eliminou a necessidade de julgamento. Ele mudou o lugar em que esse julgamento precisa acontecer.

Essa trajetória desloca a pergunta sobre se a inteligência artificial "substitui o diagnóstico médico" para um problema mais preciso: como um algoritmo aprende a reconhecer doença, quais evidências sustentam esse reconhecimento e o que muda quando uma classificação computacional entra em uma decisão clínica. É nesse ponto — onde capacidade de reconhecer padrões encontra consequência concreta para o paciente — que a promessa tecnológica começa a ser julgada pela qualidade do cuidado que efetivamente produz.

CAPÍTULO 2
QUANDO O ALGORITMO APRENDE A RECONHECER DOENÇA: DIAGNÓSTICO, IMAGEM E DETECÇÃO DE PADRÕES

Acertar o rótulo não é o mesmo que diagnosticar

Uma das imagens mais sedutoras da inteligência artificial em saúde é a de um sistema capaz de olhar para uma radiografia, uma lâmina histológica ou uma fotografia de retina e identificar uma doença que talvez passasse despercebida ao olhar humano. Essa imagem não é ficção. Em diferentes especialidades, algoritmos já conseguem reconhecer padrões com desempenho elevado em tarefas cuidadosamente definidas. O problema começa quando esse desempenho é traduzido, sem mediação, pela frase "a inteligência artificial diagnostica melhor". Entre reconhecer um padrão em um conjunto de dados e melhorar o diagnóstico de uma pessoa existe um caminho clínico muito mais longo.

Diagnóstico não é simplesmente atribuir um rótulo correto a uma imagem. É um processo que combina probabilidade, contexto, história clínica, qualidade do exame, prevalência da doença, consequências do erro e decisões que vêm depois do resultado. Uma lesão com determinada aparência pode ter significado diferente conforme idade, sintomas, fatores de risco e exames prévios. Da mesma forma, um teste que apresenta excelente desempenho em uma população selecionada pode produzir resultados muito diferentes quando utilizado em uma população na qual a doença é rara, os equipamentos mudam ou a qualidade das imagens é menos controlada.

Essa diferença ajuda a entender por que métricas de desempenho precisam ser lidas como partes de uma história, e não como veredictos. Sensibilidade informa quão bem um sistema identifica pessoas que realmente apresentam a condição de interesse; especificidade mostra sua capacidade de reconhecer quem não a apresenta. Alterar o ponto de corte pode favorecer uma delas em detrimento da outra. Em rastreamento de uma doença grave, pode ser aceitável tolerar mais falsos positivos para reduzir o risco de perder casos. Em outra situação, uma avalanche de resultados falsamente positivos pode produzir exames invasivos, ansiedade, custos e sobrecarga que anulam parte do benefício esperado.

A área sob a curva ROC, frequentemente apresentada como símbolo de excelência de um modelo, também precisa ser colocada em seu lugar. Ela resume capacidade de discriminação ao longo de diferentes limiares, mas não responde sozinha se as probabilidades fornecidas pelo modelo estão bem calibradas, se o limiar escolhido é clinicamente sensato ou se o uso daquele sistema melhora uma decisão. Um algoritmo pode ordenar corretamente pacientes de maior e menor risco e ainda assim fornecer probabilidades inadequadas para a prática. Em saúde, discriminar bem é importante; saber o que fazer com essa discriminação é outra questão.

Essa preocupação ganhou forma explícita nas orientações metodológicas recentes. O TRIPOD+AI, publicado em 2024, atualizou as recomendações de transparência para estudos de modelos de predição diagnóstica e prognóstica, incorporando questões próprias de métodos de aprendizagem de máquina e reforçando a necessidade de relatar desenvolvimento, validação, incerteza, desempenho e aplicabilidade de modo suficientemente claro para que o leitor consiga julgar a força da evidência (COLLINS et al., 2024). Em 2025, o PROBAST+AI ampliou a avaliação de qualidade, risco de viés e aplicabilidade, tratando separadamente o desenvolvimento e a avaliação de modelos e examinando participantes, fontes de dados, preditores, desfechos e análise (MOONS et al., 2025).

O que essas ferramentas metodológicas tornam evidente é que um número alto de desempenho pode esconder uma pergunta mal respondida. Quem compôs o conjunto de dados? O padrão de referência usado para definir a "verdade" diagnóstica era confiável? Imagens do mesmo paciente chegaram inadvertidamente aos conjuntos de treinamento e teste? O modelo foi avaliado em outra instituição, em outro equipamento ou em outra população? Foram analisadas diferenças entre subgrupos? Casos tecnicamente difíceis foram excluídos? Quando essas informações faltam, a precisão aparente pode refletir, em parte, condições ideais de pesquisa que não existirão depois da implantação.

O próprio conceito de padrão de referência merece atenção. Algoritmos aprendem a partir de rótulos, mas rótulos clínicos nem sempre representam uma verdade absoluta. Em radiologia e patologia, especialistas podem discordar; em doenças que dependem de evolução temporal, a confirmação pode exigir acompanhamento; em bancos retrospectivos, o "diagnóstico" pode ser extraído de prontuários com erros, omissões ou critérios heterogêneos. Se o conjunto de treinamento incorpora essas incertezas, o modelo não aprende uma doença em estado

puro. Aprende uma representação produzida pelos dados e pelas decisões humanas que deram origem aos rótulos.

Isso significa que a inteligência artificial não elimina subjetividade simplesmente porque utiliza matemática. Parte da subjetividade, da prática institucional e das desigualdades do sistema pode ser codificada nos dados. Um serviço que solicita determinado exame mais frequentemente para um grupo de pacientes produzirá um banco diferente de outro serviço com práticas distintas. Um algoritmo treinado sobre essa história pode reconhecer tanto sinais biológicos quanto marcas indiretas de como o cuidado foi organizado. O modelo é capaz de encontrar correlações que os pesquisadores nem sequer perceberam, e justamente por isso é necessário perguntar se essas correlações continuarão válidas quando o contexto mudar.

A mudança de contexto é um dos maiores testes da validade clínica. Um modelo desenvolvido em um hospital terciário pode encontrar, durante o treinamento, pacientes mais graves do que aqueles atendidos na atenção primária. Um sistema treinado em imagens de um fabricante pode sofrer alteração de desempenho em outro equipamento. Protocolos de aquisição, prevalência de doenças, composição demográfica e condutas locais podem modificar a distribuição dos dados. Esse fenômeno, frequentemente descrito como mudança de distribuição, não significa que o algoritmo "parou de funcionar" de forma misteriosa; significa que o mundo no qual ele foi aplicado deixou de ser suficientemente parecido com o mundo em que aprendeu.

Por isso, validação externa é mais do que um rito metodológico. Ela é uma tentativa de descobrir se o modelo consegue sobreviver fora do ambiente que o produziu. Mesmo assim, uma única validação externa não encerra a questão. Um sistema pode funcionar em dois centros e falhar em um terceiro; pode manter desempenho global e piorar em determinado subgrupo; pode ser adequado em uma etapa do cuidado e inadequado em outra. O diagnóstico assistido por IA precisa ser pensado como uma tecnologia situada: seu valor depende da relação entre modelo, população, tarefa e fluxo clínico.

Essa forma de pensar muda a pergunta. Em vez de perguntar simplesmente "o algoritmo acerta?", passamos a perguntar "em quem, em que condições, comparado com qual padrão e com que consequência?". É quando essa pergunta ganha concretude que alguns dos exemplos mais maduros de IA diagnóstica se tornam particularmente instrutivos.

Quando a imagem se torna terreno fértil para a inteligência artificial

As imagens médicas se tornaram um dos campos mais férteis para a inteligência artificial porque combinam três características favoráveis: grandes volumes de dados digitalizados, tarefas visualmente estruturadas e decisões que podem ser comparadas com padrões de referência relativamente bem definidos. Radiologia, oftalmologia, dermatologia e patologia digital passaram, por isso, a ocupar lugar central no desenvolvimento de sistemas capazes de localizar, classificar ou priorizar achados. A lista de dispositivos médicos habilitados por IA mantida pela FDA reflete essa presença expressiva da imagem no ecossistema regulado, ainda que a própria agência ressalte que a lista não é exaustiva (FDA, 2026a).

Mas a imagem também ensina uma lição importante: o mesmo algoritmo pode ocupar papéis muito diferentes. Em alguns cenários, atua como segunda leitura, destacando regiões suspeitas para um profissional. Em outros, reorganiza a fila e prioriza exames potencialmente urgentes. Há sistemas de triagem que decidem quais casos precisam de leitura adicional e sistemas autônomos que emitem uma conclusão sem interpretação especialista obrigatória daquele exame. Dizer apenas que "há IA no diagnóstico" esconde essas diferenças, embora elas mudem profundamente risco, responsabilidade e impacto sobre o fluxo de trabalho.

A retinopatia diabética oferece um dos exemplos mais claros de autonomia diagnóstica. No estudo pivotal de Abràmoff et al. (2018), um sistema autônomo foi avaliado prospectivamente em unidades de atenção primária para identificar retinopatia diabética em pessoas com diabetes que não tinham diagnóstico prévio da complicação. Em 900 participantes, o sistema atingiu sensibilidade de 87,2% e especificidade de 90,7% para o desfecho definido no estudo, além de conseguir produzir resultado em grande parte dos pacientes. O aspecto transformador não era apenas a classificação da fotografia de retina: era a possibilidade de realizar a avaliação no próprio ponto de cuidado, sem depender de uma interpretação oftalmológica prévia para cada exame.

Essa mudança de local importa porque acesso faz parte do valor diagnóstico. Um exame de alta precisão disponível apenas em um serviço distante pode produzir menos benefício populacional do que um exame suficientemente preciso incorporado ao lugar onde as pessoas já recebem cuidado. Ao levar o rastreamento para a atenção primária, a IA modifica o caminho do paciente: quem apresenta resultado compatível com doença é encaminhado, enquanto quem não apresenta pode concluir a etapa de

rastreamento sem uma consulta especializada apenas para leitura da imagem. Nesse caso, a tecnologia não está somente "imitando um oftalmologista"; está redesenhando uma etapa do sistema assistencial.

A experiência acumulada depois das primeiras autorizações também mostra por que o estudo pivotal não deve ser confundido com ponto final. A meta-análise de Wang et al. (2026), restrita a sistemas de aprendizagem profunda aprovados por autoridades regulatórias para detecção de retinopatia diabética, reuniu 82 estudos e 887.244 exames de 25 dispositivos. O desempenho agrupado foi elevado, mas variou conforme limiar de doença, renda do país, qualidade e gradabilidade das imagens, dilatação pupilar, padrão de referência e critérios diagnósticos. Apenas uma parte dos estudos utilizou conjuntos de teste plenamente externos. A conclusão mais importante talvez não seja o número agregado, e sim a demonstração de que desempenho real continua dependente do contexto mesmo depois da aprovação regulatória.

Mamografia oferece outra história, desta vez centrada na colaboração entre algoritmo e radiologistas. Durante anos, muitos estudos retrospectivos sugeriram que modelos poderiam detectar câncer de mama com desempenho comparável ao de leitores humanos. Esse tipo de evidência era relevante, mas deixava aberta uma pergunta decisiva: o que acontece quando a IA participa de um programa de rastreamento de verdade, com mulheres reais, radiologistas reais e consequências reais para chamadas adicionais, biópsias e cânceres diagnosticados entre uma rodada e outra?

O ensaio MASAI, conduzido dentro do programa sueco de rastreamento mamográfico, foi desenhado precisamente para aproximar a avaliação dessa realidade. Na análise publicada em 2025, o fluxo apoiado por IA detectou mais cânceres do que a dupla leitura padrão e reduziu substancialmente a carga de leitura, sem aumentar os falsos positivos; o aumento de detecção foi concentrado em tumores invasivos pequenos, sem comprometimento linfonodal, além de determinados carcinomas in situ de maior grau (HERNSTRÖM et al., 2025). O resultado já era mais informativo do que comparar um algoritmo isolado a radiologistas em um banco retrospectivo, porque avaliava um procedimento assistencial completo.

A análise de seguimento publicada no The Lancet em 2026 tornou a história ainda mais interessante. Com a população do estudo acompanhada entre as rodadas de rastreamento, Gommers et al. (2026) observaram que o protocolo apoiado por IA

foi não inferior ao padrão quanto à ocorrência de cânceres de intervalo, apresentou sensibilidade de 80,5% contra 73,8% na dupla leitura convencional e manteve especificidade de 98,5% em ambos os grupos. A taxa de câncer de intervalo foi 12% menor no grupo com IA, embora a estimativa tenha intervalo de confiança compatível com incerteza sobre a magnitude do efeito. Isso é cientificamente mais relevante do que dizer apenas que "a IA detectou mais": começa a responder se os tumores adicionais encontrados alteram o desempenho do rastreamento ao longo do tempo.

Mesmo esses resultados não autorizam afirmar que o uso de IA em mamografia já demonstrou redução de mortalidade por câncer de mama. Desfechos intermediários, como sensibilidade, câncer de intervalo, estágio e características dos tumores detectados, são clinicamente importantes, mas não equivalem automaticamente a benefício final em sobrevida. Essa diferença ilustra uma regra que acompanhará todo este livro: quanto maior a distância entre o desfecho medido e aquilo que realmente importa ao paciente, maior deve ser a prudência na interpretação.

A experiência alemã acrescenta outra camada: implementação em grande escala fora de um ensaio randomizado. O estudo PRAIM integrou um sistema de IA ao programa populacional de mamografia em 12 centros e analisou mais de 460 mil mulheres, com 119 radiologistas e equipamentos de diferentes fabricantes. A leitura apoiada por IA apresentou maior taxa de detecção de câncer com taxa de reconvocação semelhante ou menor em comparação ao fluxo sem IA, em um desenho prospectivo observacional incorporado à rotina (EISEMANN et al., 2025). O tamanho e a diversidade operacional aumentam a relevância externa, mas o desenho não randomizado e o envolvimento do fornecedor no financiamento e em partes do estudo precisam permanecer visíveis na interpretação.

Esses exemplos mostram por que "IA versus médico" costuma ser uma pergunta pobre. Em retinopatia diabética, a tecnologia pode funcionar autonomamente em um ponto do percurso. Em mamografia, o valor emerge de uma arquitetura em que o algoritmo estratifica exames, apoia a leitura e redistribui trabalho humano. Em radiografias de tórax, estudos de leitores também sugerem que a assistência algorítmica pode melhorar dimensões de sensibilidade e eficiência em tarefas específicas, embora o efeito dependa da tarefa, da experiência do usuário e do modo como a recomendação é apresentada (AHN et al., 2022). O objeto clínico relevante não é sempre o algoritmo isolado; muitas vezes é o sistema humano-tecnológico.

Isso também muda o significado do erro. Um falso negativo produzido por um sistema autônomo pode encerrar uma etapa de rastreamento sem avaliação especializada. Em um sistema de segunda leitura, o profissional ainda pode contrariar a recomendação. Já um falso positivo usado para priorizar uma fila pode aumentar exames urgentes sem necessariamente produzir um diagnóstico final incorreto. O mesmo valor de sensibilidade pode ter consequências distintas conforme a posição ocupada pela IA no fluxo.

A arquitetura do cuidado, portanto, precisa fazer parte da avaliação diagnóstica. Quem recebe o resultado? Há possibilidade de revisão? O sistema expressa incerteza ou apenas fornece uma classe? O usuário sabe em quais situações o modelo não deveria ser usado? Como são tratadas imagens tecnicamente inadequadas? Há mecanismo para encaminhar casos inconclusivos? Essas perguntas parecem operacionais, mas são também perguntas de segurança.

Da curva de desempenho ao paciente real

A passagem do laboratório para a clínica exige uma mudança de escala e de linguagem. No laboratório, é possível perguntar se o modelo diferencia corretamente duas classes. Na clínica, é preciso observar o que profissionais fazem com a recomendação, como pacientes percorrem o sistema, quais erros aparecem e se o novo fluxo produz benefício suficiente para justificar seus custos e riscos. É nessa transição que muitos modelos promissores deixam de ser apenas problemas de ciência de dados e se tornam intervenções em saúde.

O DECIDE-AI foi desenvolvido justamente para esse estágio inicial de avaliação ao vivo. A diretriz enfatiza que sistemas de suporte baseados em IA devem ser estudados em condições clínicas reais, considerando não apenas desempenho técnico, mas utilidade, segurança, fatores humanos e integração no ambiente de uso (VASEY et al., 2022). A proposta é particularmente importante porque o comportamento de um profissional diante de uma recomendação não pode ser deduzido de uma curva ROC. Pessoas podem confiar demais, ignorar alertas, modificar padrões de busca visual ou desenvolver estratégias próprias de interação com a ferramenta.

A confiança excessiva é um risco conhecido. Se a IA adquire reputação de alta precisão, o usuário pode reduzir a busca por evidências contrárias e aceitar uma saída errada com menos questionamento. O efeito oposto também existe: sistemas que produzem muitos alertas irrelevantes podem ser progressivamente ignorados, inclusive quando um alerta importante aparece. Entre dependência e rejeição existe um

problema de calibração humana: o profissional precisa aprender quando a ferramenta costuma ajudar e quando seus limites exigem maior vigilância.

Essa aprendizagem não ocorre uma única vez. Modelos, interfaces e populações mudam. Um hospital pode trocar equipamentos, alterar protocolos de imagem ou ampliar atendimento para grupos que estavam pouco representados no desenvolvimento. A epidemiologia de uma doença também muda. Atualizações do software podem modificar comportamento. Por isso, implantação não é o momento em que a avaliação termina; é o momento em que começa uma nova fase de evidência.

Monitoramento pós-implementação precisa observar mais do que falhas técnicas. Uma queda de sensibilidade pode ser sinal de mudança de distribuição; aumento de falsos positivos pode resultar de alteração no equipamento; desempenho estável no conjunto pode esconder piora em subgrupos. Também é necessário acompanhar o efeito sobre tempos de trabalho, filas, solicitações adicionais, encaminhamentos e acesso. Um sistema pode ser tecnicamente preciso e ainda produzir gargalos em outra parte do serviço.

A avaliação por subgrupos merece lugar central porque desempenho médio pode esconder desigualdade. Se um algoritmo é excelente na população mais numerosa e claramente inferior em um grupo menor, a média pode continuar atraente. Sexo, idade, raça e etnia, tonalidade de pele, presença de comorbidades, fabricante do equipamento e local de atendimento podem funcionar como marcadores relevantes conforme a aplicação. A questão não é exigir desempenho idêntico em qualquer circunstância, mas tornar diferenças visíveis e decidir se são clinicamente aceitáveis.

Esse princípio está ligado à representatividade dos dados, mas não se resolve apenas aumentando diversidade no treinamento. Mesmo um banco aparentemente diverso pode não reproduzir a combinação de prevalência, fluxo, critérios de encaminhamento e qualidade de exame encontrada em determinado serviço. A melhor proteção continua sendo avaliar o modelo no contexto em que se pretende utilizá-lo e manter vigilância depois da implantação.

A aprovação regulatória, por isso, deve ser entendida como marco importante de um processo, não como certificado permanente de benefício. Ela indica que um produto atendeu critérios específicos para determinada finalidade, versão e população de uso. Não significa que qualquer uso semelhante esteja automaticamente validado nem que o desempenho permanecerá constante em todas as instituições. Sistemas de

IA são particularmente sensíveis a esse problema porque atualizações podem alterar algoritmos e porque a própria prática clínica evolui.

A discussão sobre calibração volta a aparecer nesse ponto. Para alguns sistemas, a saída não é apenas "doença" ou "não doença", mas uma probabilidade. Uma probabilidade de 20% só é clinicamente significativa se, entre pessoas com previsão semelhante, a frequência observada do desfecho estiver aproximadamente alinhada e se o ponto de corte escolhido fizer sentido no contexto. Modelos mal calibrados podem induzir decisões desproporcionais, mesmo quando discriminam relativamente bem quem tem maior e menor risco. Essa é uma das razões pelas quais diretrizes de modelagem atuais insistem que desempenho não pode ser resumido a uma única métrica.

Também é necessário definir o comparador correto. Se a prática padrão é dupla leitura por radiologistas, comparar IA isolada a um único leitor pode responder a uma pergunta interessante, mas não à pergunta do sistema de saúde. Se o objetivo é ampliar rastreamento para pessoas que hoje não conseguem acesso, o comparador pode ser a ausência de exame, não um especialista disponível em condições ideais. O valor de uma tecnologia depende da alternativa real que ela pretende substituir ou complementar.

Essa observação é especialmente importante em países com desigualdade de acesso. Um sistema autônomo de rastreamento pode ampliar cobertura em regiões com escassez de especialistas, mas precisa funcionar com equipamentos, conectividade e perfis populacionais compatíveis com aquele território. Se depender de infraestrutura indisponível ou encaminhar grande volume de falsos positivos para uma rede incapaz de absorvê-los, a inovação pode deslocar o gargalo em vez de resolvê-lo. Diagnóstico é parte de uma cadeia; melhorar um elo não garante melhoria do percurso inteiro.

A mesma prudência vale para custos. Reduzir carga de leitura pode liberar tempo especializado, mas aquisição, licenciamento, integração, manutenção, treinamento, monitoramento e resposta a incidentes também consomem recursos. Benefício clínico e eficiência econômica não são sinônimos, embora possam se reforçar. Uma tecnologia com ganho diagnóstico pequeno pode ser valiosa se permitir cobrir uma população antes desassistida; outra, com excelente desempenho técnico, pode ter baixo valor se duplicar trabalho ou exigir infraestrutura desproporcional.

Por isso, a pergunta "a IA diagnostica melhor?" precisa ser substituída por um conjunto de perguntas mais difíceis: melhora a sensibilidade sem produzir dano inaceitável por falsos positivos? Mantém desempenho em diferentes grupos e ambientes? Reduz ou aumenta carga de trabalho? O profissional consegue interpretar e contestar a saída? O sistema falha de forma detectável? Existe caminho para casos inconclusivos? O resultado chega a quem precisa no tempo correto? E, sobretudo, a mudança modifica desfechos relevantes para o paciente ou apenas melhora uma métrica intermediária?

Essas perguntas não diminuem a importância dos avanços alcançados. Pelo contrário, ajudam a identificar onde a inteligência artificial já ultrapassou a fase de demonstração técnica. O rastreamento autônomo de retinopatia diabética e os estudos prospectivos de mamografia mostram que é possível avaliar IA dentro de fluxos reais e produzir evidência mais próxima da decisão assistencial. Ao mesmo tempo, eles mostram que o valor aparece quando tecnologia, população, profissionais e organização do cuidado são estudados juntos.

Reconhecer doença, entretanto, é apenas uma parte do trabalho clínico. Depois que um sistema estima a presença de uma condição ou identifica um padrão de risco, surge a pergunta que realmente organiza a conduta: o que fazer com essa informação? Escolher exames, priorizar pacientes, estimar prognóstico, comparar tratamentos e adaptar decisões às características individuais desloca a inteligência artificial de uma função predominantemente perceptiva para uma função decisória. A partir daí, o valor clínico já não depende apenas de acertar uma classificação, mas de como previsão, incerteza e contexto se transformam em escolha.

CAPÍTULO 3
DA PREVISÃO À DECISÃO: SUPORTE CLÍNICO, PROGNÓSTICO E MEDICINA PERSONALIZADA

Prever o futuro não é escolher o caminho

O reconhecimento de uma doença responde a uma pergunta sobre o presente: há ou não há um padrão compatível com determinada condição? O prognóstico desloca o tempo. Ele pergunta o que pode acontecer nas próximas horas, dias, meses ou anos e com que probabilidade. Essa mudança parece pequena, mas altera profundamente o lugar da inteligência artificial no cuidado. Um sistema que identifica uma pneumonia em uma radiografia está contribuindo para interpretar um dado existente; um sistema que estima deterioração clínica, readmissão, mortalidade ou falência de órgão começa a disputar atenção sobre eventos que ainda não ocorreram. A utilidade passa a depender não apenas de acertar, mas de antecipar algo em tempo suficiente para que uma conduta relevante ainda seja possível.

A medicina sempre trabalhou com previsão. Escalas de risco, estadiamentos, escores prognósticos e julgamento clínico são formas de organizar sinais presentes para estimar futuros possíveis. O que o aprendizado de máquina acrescentou foi a capacidade de combinar grande número de variáveis, explorar relações não lineares e atualizar estimativas à medida que novos dados entram no prontuário. Obermeyer e Emanuel (2016) já apontavam que uma das aplicações mais promissoras do aprendizado de máquina seria justamente o prognóstico. Dois anos depois, Rajkomar et al. (2018) mostraram que redes neurais aplicadas a registros eletrônicos brutos podiam produzir previsões para mortalidade hospitalar, readmissão e permanência prolongada sem depender da mesma engenharia manual de variáveis exigida por modelos tradicionais.

Esse avanço muda a escala do problema, mas não muda sua natureza fundamental. Toda previsão depende de três escolhas: qual evento se deseja antecipar, em qual horizonte de tempo e para qual finalidade. Estimar risco de morte em doze meses pode ser útil para planejamento de cuidados em determinado contexto e completamente inútil para decidir qual paciente precisa ser revisto nos próximos vinte minutos. Da mesma forma, um modelo que prevê readmissão pode ser estatisticamente competente e ainda oferecer pouco valor se o serviço não tiver uma

intervenção capaz de modificar esse risco. Uma probabilidade só se transforma em informação clínica quando existe uma ação plausível associada a ela.

A previsão contínua de lesão renal aguda ilustra bem essa tensão. Tomašev et al. (2019) desenvolveram um sistema de aprendizado profundo usando registros de mais de setecentos mil adultos atendidos no sistema de saúde dos veteranos dos Estados Unidos. O modelo procurava antecipar episódios de lesão renal aguda com até 48 horas de antecedência e identificava grande parte dos casos mais graves que posteriormente necessitariam de diálise. Ao mesmo tempo, havia falsos alertas: para cada alerta verdadeiro, o estudo relatou aproximadamente dois falsos. O resultado é tecnicamente impressionante e clinicamente provocador, mas não responde sozinho ao que deveria acontecer depois de cada aviso.

Esse é o ponto em que previsão e decisão começam a se separar. Se um alerta de risco leva a revisão de medicamentos nefrotóxicos, avaliação de volume, repetição de exames ou maior vigilância, pode criar uma janela de prevenção. Mas cada ação também possui custo, risco e demanda de trabalho. Aumentar coletas laboratoriais em centenas de pacientes pode produzir desconforto, consumo de recursos e novas decisões baseadas em alterações sem relevância. Suspender um medicamento por receio de lesão renal pode comprometer o tratamento de outra condição. Antecipar um problema não elimina o conflito entre possíveis benefícios e danos.

Por isso, o valor de um modelo prognóstico depende de sua capacidade de produzir risco acionável. A expressão é importante porque distingue informação interessante de informação útil. Um risco é acionável quando chega à pessoa certa, no momento certo, com antecedência suficiente, acompanhado de uma possibilidade de resposta e inserido em um fluxo que consegue absorver essa resposta. Um sistema que prevê deterioração horas antes do evento, mas gera o aviso em uma tela que ninguém consulta, tem desempenho computacional sem desempenho assistencial.

Também existe uma diferença entre prever um desfecho e explicar por que ele ocorrerá. Modelos treinados em registros eletrônicos podem descobrir combinações altamente preditivas que funcionam como marcadores de gravidade, mas não necessariamente como causas modificáveis. Uma variável pode ajudar a identificar quem morrerá sem oferecer qualquer pista sobre qual tratamento evitaria a morte. Esse é um erro conceitual perigoso: transformar correlação prognóstica em justificativa terapêutica.

A distinção fica ainda mais clara quando se compara risco com benefício de tratamento. Imagine dois pacientes com elevada probabilidade de um evento cardiovascular. O primeiro pode ter grande possibilidade de benefício com determinada intervenção; o segundo pode apresentar contraindicações, risco elevado de efeitos adversos ou características que reduzem a efetividade esperada. Um modelo que estima apenas a probabilidade do evento não conhece necessariamente a diferença entre esses cenários. Ele informa quem está em maior risco, mas não demonstra quem mais se beneficiará da conduta.

Kent, Steyerberg e van Klaveren (2018) tratam essa questão no campo da medicina personalizada ao distinguir previsão de risco e heterogeneidade do efeito de tratamento. Resultados médios de ensaios clínicos são obtidos em grupos, enquanto decisões são tomadas para indivíduos. A tentativa de personalizar cuidado exige estimar como benefício e dano podem variar entre pessoas, e não apenas identificar quem possui maior probabilidade do desfecho. Em alguns tratamentos, risco basal alto está associado a maior benefício absoluto; em outros, modificadores de efeito podem alterar a resposta de formas que um simples escore prognóstico não captura.

Essa diferença estabelece um limite para a promessa da chamada medicina de precisão. Ter mais dados sobre uma pessoa não garante automaticamente que saibamos qual decisão é melhor para ela. Genômica, exames, imagens, dados de sensores, histórico longitudinal e informações sociais podem refinar uma representação do paciente. Ainda assim, transformar essa representação em escolha terapêutica requer evidência sobre o que acontece sob alternativas diferentes. A pergunta deixa de ser "o que provavelmente acontecerá?" e passa a ser "o que provavelmente acontecerá se fizermos A, em comparação com B?".

A segunda pergunta é causal. Ela não pode ser respondida apenas observando que determinado tratamento foi mais frequente entre pacientes que tiveram melhor evolução, porque escolhas terapêuticas na prática não são aleatórias. Médicos tendem a prescrever tratamentos diferentes conforme gravidade, comorbidades, idade, preferência e inúmeras variáveis que também influenciam o desfecho. Um algoritmo treinado sobre decisões históricas pode aprender tanto a medicina quanto os vieses, desigualdades e limitações existentes no sistema que gerou os dados.

É por isso que a passagem de prognóstico para recomendação exige uma mudança de padrão de evidência. Para prever, pode ser suficiente demonstrar que o modelo estima bem um desfecho em populações relevantes. Para recomendar tratamento, é

necessário sustentar que seguir a recomendação melhora a relação entre benefícios e danos em comparação com alternativas. O que se exige da tecnologia cresce porque cresce o poder que seu resultado exerce sobre o cuidado.

Essa diferença também ajuda a compreender por que suporte à decisão clínica não deve ser pensado como uma máquina que "sabe a resposta". Shortliffe e Sepúlveda (2018) observam que sistemas de suporte podem integrar informações e auxiliar profissionais em situações complexas, mas precisam ser concebidos dentro da interação entre ferramenta e usuário. A decisão emerge dessa interação, não apenas da saída do algoritmo. Quando essa saída entra em uma rotina clínica concreta, fatores humanos passam a ser tão relevantes quanto a estatística que produziu a previsão.

Quando o alerta entra na rotina

Poucas situações mostram melhor a diferença entre modelo e sistema de cuidado do que a sepse. Reconhecê-la precocemente é difícil porque sinais iniciais podem ser inespecíficos, a evolução pode ser rápida e atrasos no tratamento têm consequências importantes. Por isso, modelos de alerta tornaram-se um terreno atraente para inteligência artificial e aprendizado de máquina. A promessa é intuitiva: reunir continuamente dados do prontuário, identificar padrões que precedem deterioração e chamar a atenção da equipe antes que o quadro se torne evidente.

A história recente, entretanto, mostra por que desempenho publicado não deve ser confundido com segurança operacional. Wong et al. (2021) realizaram validação externa de um modelo proprietário de sepse amplamente implementado em prontuários eletrônicos nos Estados Unidos. No hospital estudado, o sistema apresentou desempenho substancialmente inferior ao que seria desejável para um alerta de alto impacto, com baixa sensibilidade em um ponto de corte utilizado na prática e grande quantidade de alertas. O estudo se tornou emblemático não porque prove que modelos de sepse são inúteis, mas porque mostra que adoção em larga escala pode preceder validação independente suficiente.

Um alerta ruim cria dois tipos de problema. Se deixa de identificar pacientes que realmente estão deteriorando, pode produzir falsa tranquilidade. Se sinaliza pessoas demais, compete com dezenas de outros avisos, exames e prioridades pela atenção dos profissionais. A sensibilidade, portanto, não pode ser maximizada isoladamente. Cada aumento de detecção pode trazer uma carga adicional de falsos positivos, e essa carga não é apenas estatística. Ela aparece como cliques, avaliações, interrupções e tempo.

O fenômeno é conhecido há muito tempo nos sistemas de suporte à decisão. Goddard, Roudsari e Wyatt (2012), em revisão sistemática sobre viés de automação, mostraram que usuários podem seguir recomendações incorretas de sistemas automatizados e que fatores como carga de trabalho, complexidade e pressão de tempo influenciam essa dependência. O risco não é simplesmente "confiar em tecnologia". É calibrar mal a confiança: aceitar um resultado quando seria necessário contestá-lo ou ignorar o sistema quando ele tem informação valiosa.

Há ainda o problema inverso da confiança excessiva: a fadiga de alertas. Quando grande proporção dos avisos é irrelevante, profissionais aprendem que interromper o trabalho para examiná-los raramente vale a pena. O comportamento é racional no curto prazo, mas perigoso quando um alerta importante se mistura ao ruído. Nesse cenário, um sistema inicialmente criado para aumentar atenção termina treinando a equipe a desconsiderá-lo.

A experiência do TREWS, um sistema de alerta para sepse estudado em cinco hospitais, oferece um contraste útil porque a avaliação não se limitou à curva de desempenho. Henry et al. (2022) analisaram como profissionais realmente interagiam com os avisos ao longo de dois anos. Entre os casos de sepse identificados retrospectivamente, o sistema sinalizou grande proporção; a maioria dos alertas foi revisada por médicos ou profissionais de prática avançada, mas apenas parte deles foi confirmada como suspeita de sepse. Quando um alerta verdadeiro era confirmado em até três horas, o primeiro pedido de antibiótico ocorria mais cedo do que entre casos em que a confirmação acontecia depois ou não acontecia.

O mesmo grupo acompanhou resultados após a implementação. Adams et al. (2022) conduziram estudo prospectivo multicêntrico com o TREWS e observaram associação entre confirmação rápida de alertas e menor mortalidade entre pacientes com sepse, além de tratamento mais oportuno. O desenho não equivale a um ensaio randomizado que prove causalmente que o sistema sozinho produziu a melhora, e essa distinção precisa ser preservada. Ainda assim, o estudo avança em relação à avaliação puramente retrospectiva porque observa comportamento de profissionais, tempo de resposta e desfechos dentro de um fluxo real.

O contraste entre os dois exemplos é didático. Não existe uma categoria homogênea chamada "algoritmo de sepse". Ferramentas podem diferir em dados, atualização temporal, limiares, interface, treinamento da equipe, integração ao prontuário e estratégia de implementação. Mesmo modelos com desempenho

semelhante podem gerar resultados diferentes se um estiver incorporado a um processo claro de resposta e o outro apenas acrescentar mais um aviso à tela.

Isso explica por que adoção também é um resultado clínico intermediário. Um sistema que ninguém utiliza não pode melhorar cuidado; um sistema que todos obedecem sem crítica também pode ser perigoso. O objetivo é uma relação calibrada, na qual o profissional sabe quando confiar, quando revisar, quando discordar e o que fazer diante de incerteza. Esse equilíbrio depende de design, treinamento, transparência e experiência acumulada.

A atualização da orientação da Food and Drug Administration sobre software de suporte à decisão clínica, publicada em janeiro de 2026, torna essa discussão especialmente atual. A FDA diferencia determinadas funções de CDS que podem ficar fora da definição regulatória de dispositivo médico de outras funções de software que continuam sujeitas às políticas de dispositivos. Entre os elementos centrais está a possibilidade de o profissional de saúde revisar de forma independente a base da recomendação, em vez de depender cegamente de uma saída que não consegue avaliar (FDA, 2026b). A fronteira regulatória, portanto, dialoga com uma questão clínica: quanto mais determinante é a recomendação e menos verificável é sua base, maior é a preocupação com dependência do sistema.

No Brasil, a Resolução CFM nº 2.454/2026, publicada em fevereiro e com entrada em vigor prevista 180 dias após sua publicação, segue lógica compatível ao estabelecer a inteligência artificial como instrumento de apoio e manter sob responsabilidade médica a decisão diagnóstica, terapêutica e prognóstica. O texto também exige julgamento crítico do resultado e preserva ao profissional a possibilidade de recusar tecnologias sem validação científica ou certificação pertinente (CFM, 2026). A regra não transforma o médico em auditor técnico capaz de reconstruir qualquer rede neural; ela reforça que o uso clínico não retira do profissional a obrigação de avaliar contexto, coerência e limites.

Essa responsabilidade, contudo, não pode servir para deslocar todo risco do sistema para a pessoa que está diante da tela. Se uma instituição adquire uma ferramenta opaca, configura alertas inadequadamente, não monitora desempenho e não oferece treinamento, dizer ao médico que ele possui a palavra final não resolve a falha de governança. Segurança depende de responsabilidades distribuídas entre desenvolvedores, reguladores, instituições e profissionais. O julgamento humano é

indispensável, mas precisa ser apoiado por informação suficiente e por um ambiente que permita discordar.

A interface participa dessa segurança. Mostrar apenas um número pode induzir interpretação diferente de apresentar tendência temporal, incerteza e variáveis que contribuíram para o alerta. Explicações, porém, não devem ser confundidas com prova de causalidade. Saber que creatinina, pressão arterial e uso de determinado medicamento contribuíram para a previsão não significa que modificar cada um desses elementos alterará o desfecho na proporção sugerida pelo modelo. Explicar o que influenciou a previsão e justificar o que deve ser feito são tarefas distintas.

Também importa o momento da recomendação. Um alerta precoce demais pode gerar muitos falsos positivos porque a informação disponível ainda é pouco específica. Um alerta tardio pode ser mais preciso, mas perder a janela em que a intervenção faria diferença. O ponto ótimo não é uma propriedade fixa do algoritmo; depende da urgência do problema, do custo de errar e da capacidade do serviço de responder. Em uma emergência, minutos podem importar. Em prevenção de longo prazo, talvez seja mais valioso reduzir falsos positivos e discutir alternativas com calma.

Essa dimensão temporal é uma das razões pelas quais suporte à decisão não pode ser avaliado apenas como software. Ele é uma intervenção sociotécnica. Modifica fluxo, atenção, responsabilidades e comunicação. Ao inserir uma recomendação automatizada, a organização muda também a arquitetura da escolha: algo passa a ser destacado, priorizado ou sugerido antes que o profissional formule sua própria hipótese. A tecnologia não precisa substituir a decisão para influenciá-la profundamente.

Quando esse poder de influência se torna explícito, chega-se ao ponto mais ambicioso da inteligência artificial clínica: não apenas indicar quem está em risco, mas sugerir qual tratamento, dose, sequência ou estratégia poderia produzir o melhor resultado para aquela pessoa. É nessa promessa de personalização que as fronteiras entre previsão, evidência causal e julgamento clínico ficam mais estreitas.

Personalizar não é automatizar a escolha

A medicina personalizada costuma ser descrita como a capacidade de oferecer o tratamento certo ao paciente certo no momento certo. A frase é atraente porque condensa um objetivo antigo da clínica: reconhecer que pessoas com o mesmo diagnóstico podem responder de maneiras diferentes. A inteligência artificial parece

ampliar esse projeto ao combinar sinais clínicos, exames, história, genética e dados longitudinais em escala impossível de processar manualmente. O risco é imaginar que maior resolução de dados produz, por si só, maior certeza sobre a conduta.

Para escolher entre tratamentos, o modelo precisa lidar com um problema contrafactual. Um paciente recebe A ou B, nunca os dois caminhos completos ao mesmo tempo. O resultado que teria ocorrido sob a alternativa não escolhida permanece invisível. Ensaios randomizados existem justamente para tornar comparações causais mais confiáveis em grupos. A personalização tenta avançar um passo: estimar como o efeito relativo ou absoluto varia entre indivíduos ou subgrupos. Essa tarefa é muito mais difícil do que predizer um único desfecho.

Kent, Steyerberg e van Klaveren (2018) mostram que uma forma de avançar é modelar heterogeneidade de efeito de tratamento. Pacientes com risco basal diferente podem obter benefício absoluto diferente mesmo quando o efeito relativo do tratamento é semelhante. Em outras situações, características específicas modificam a resposta ao tratamento. Modelos podem ajudar a combinar essas informações, mas a validade depende do desenho dos estudos que fornecem os dados e da forma como o efeito é estimado. Não há personalização confiável se o modelo aprende apenas padrões de prescrição histórica confundidos com gravidade e acesso.

O problema aparece com força em abordagens de aprendizado por reforço. Em 2018, Komorowski et al. apresentaram o chamado Artificial Intelligence Clinician, um agente treinado com grandes bases de pacientes com sepse para aprender estratégias de administração de fluidos e vasopressores. No conjunto de validação, a mortalidade observada era menor quando decisões clínicas históricas se aproximavam das recomendações do agente. O estudo abriu uma linha de investigação importante ao tratar cuidado intensivo como sequência de decisões, mas também ilustra por que resultados retrospectivos devem ser interpretados com prudência: concordância entre política aprendida e melhores desfechos não equivale a demonstrar que impor aquela política a novos pacientes reduzirá mortalidade.

Modelos de tratamento aprendidos em dados observacionais enfrentam um paradoxo. Eles tentam aprender com escolhas humanas e, ao mesmo tempo, superar essas escolhas. Se os melhores médicos tomavam decisões diferentes justamente porque percebiam aspectos do paciente que não foram registrados, o algoritmo pode interpretar a associação de forma errada. Se determinados grupos historicamente receberam menos acesso a uma terapia, o sistema pode transformar desigualdade

passada em recomendação futura. A qualidade do dado não se resume a quantidade; inclui aquilo que foi medido, aquilo que ficou ausente e por que determinadas decisões foram tomadas.

É por isso que medicina personalizada precisa de medicina baseada em evidências, e não de uma substituição dela. Inteligência artificial pode ajudar a identificar padrões de heterogeneidade, construir hipóteses, estratificar risco e selecionar grupos em que determinado tratamento merece ser estudado. Mas recomendações que mudam conduta precisam ser comparadas com alternativas em desenhos capazes de estimar benefício e dano. Quanto mais específica e automática é a recomendação, mais importante se torna demonstrar utilidade clínica.

A personalização também não pode ser reduzida a biomarcadores. O tratamento considerado ótimo matematicamente pode não ser o tratamento que melhor corresponde aos objetivos daquela pessoa. Um paciente pode valorizar tempo fora do hospital mais do que pequeno ganho de sobrevida; outro pode aceitar efeitos adversos maiores para maximizar chance de resposta; alguém pode priorizar preservar cognição, fertilidade, autonomia ou capacidade funcional. Esses valores não são ruído a ser eliminado por um modelo. São parte constitutiva da decisão.

Nesse sentido, a inteligência artificial pode ampliar informação sem eliminar deliberação. Um sistema pode apresentar probabilidades sob diferentes cenários, organizar evidências, estimar risco de toxicidade e identificar pacientes semelhantes. A conversa clínica continua necessária para transformar essas estimativas em escolha. A Organização Mundial da Saúde (2021) coloca autonomia humana entre os princípios centrais de governança da IA em saúde justamente porque decisões de saúde envolvem direitos, preferências e consequências que não podem ser delegados a uma função de otimização.

Há ainda uma questão de incerteza. Modelos frequentemente apresentam uma saída numérica com aparência de precisão: 17%, 42%, 73%. O número pode esconder incerteza estatística, variação entre contextos e lacunas de dados. Em medicina personalizada, comunicar incerteza é especialmente importante porque quanto mais individual parece a estimativa, maior a tentação de tratá-la como destino. Personalização responsável não promete certeza individual; procura tornar a incerteza mais informativa.

Essa diferença muda o desenho desejável de um sistema de suporte. Em vez de declarar “use o tratamento A”, uma ferramenta pode mostrar risco basal, benefício

esperado sob alternativas, possíveis danos, qualidade da evidência e grau de incerteza. Pode também indicar quando o paciente está fora do domínio em que o modelo foi validado. A interface deixa de funcionar como autoridade e passa a funcionar como estrutura para raciocínio.

Mesmo essa arquitetura exige alfabetização. Profissionais precisam compreender conceitos como calibração, probabilidade, limiares e generalização, não para se tornarem cientistas de dados, mas para reconhecer quando uma recomendação merece confiança. Desenvolvedores, por sua vez, precisam compreender o fluxo clínico e as consequências reais de um falso positivo, de uma recomendação tardia ou de uma interface ambígua. A colaboração humano-máquina começa antes do momento da decisão; ela começa no desenvolvimento e na avaliação da ferramenta.

As evidências atuais sugerem que o campo ainda está nessa transição. Há modelos capazes de antecipar deterioração, sistemas de alerta incorporados a hospitais e experiências de personalização cada vez mais sofisticadas. Ao mesmo tempo, avaliações prospectivas e estudos de impacto continuam muito menos numerosos do que trabalhos de desenvolvimento e validação técnica. Essa assimetria é um lembrete importante: a velocidade com que se produz um modelo é maior do que a velocidade necessária para demonstrar que ele melhora cuidado.

O futuro mais plausível, portanto, não é uma clínica em que algoritmos escolhem enquanto profissionais executam. É uma clínica em que diferentes formas de inteligência se distribuem tarefas. Máquinas podem vigiar séries temporais, integrar milhares de variáveis e estimar riscos repetidamente. Profissionais podem reconhecer contexto, perceber informação ausente, discutir valores e assumir responsabilidade por escolhas. A vantagem aparece quando uma capacidade corrige a limitação da outra, e não quando uma é tratada como substituta universal.

Essa complementaridade, porém, não é automática. Um sistema pode ampliar raciocínio ou estreitá-lo. Pode chamar atenção para algo ignorado ou criar ancoragem em uma hipótese equivocada. Pode reduzir variação injustificada ou reproduzir padrões históricos de exclusão. Pode oferecer tempo ao profissional ou ocupar ainda mais atenção. Tudo depende de como foi treinado, validado, inserido e monitorado.

A passagem da previsão para a decisão, contudo, não encerra o problema. Grande parte das situações ainda pode ser representada como um episódio: o paciente é avaliado, o modelo calcula um risco e alguém decide. Doenças crônicas, recuperação pós-alta e deteriorações fora do hospital, porém, não obedecem a esse recorte. Quando

sensores, prontuários e dispositivos passam a atualizar dados repetidamente, a inteligência artificial deixa de ser apenas consultada em um momento e passa a acompanhar trajetórias. O desafio muda: transformar monitoramento contínuo em cuidado útil sem converter cada oscilação em alerta, cada alerta em intervenção e cada ausência de dado em segurança presumida. Nesse cenário, decidir bem também significa saber quando agir, quando esperar e como interpretar sinais que se acumulam ao longo do tempo.

CAPÍTULO 4
CUIDADO CONTÍNUO E SISTEMAS INTELIGENTES: MONITORAMENTO, DISPOSITIVOS, ATENÇÃO REMOTA E SAÚDE POPULACIONAL

Quando o dado deixa de ser uma fotografia

Grande parte da medicina foi organizada em torno de encontros. O paciente chega à consulta, ao pronto atendimento ou ao hospital; naquele intervalo são obtidos sinais, exames e relatos; a equipe interpreta uma fotografia clínica e decide o que fazer até a próxima avaliação. Doenças crônicas, recuperação após alta e deteriorações silenciosas, porém, não respeitam essa arquitetura. Pressão arterial varia entre manhã e noite, arritmias podem desaparecer antes de um eletrocardiograma, sintomas oscilam e alterações fisiológicas podem começar dias antes de se tornarem suficientemente intensas para levar alguém de volta ao serviço. O monitoramento remoto tenta preencher justamente esse espaço entre os encontros.

A inteligência artificial modifica esse cenário porque torna computacionalmente viável examinar séries temporais extensas, reconhecer combinações de mudanças e atualizar estimativas repetidamente. O objeto deixa de ser apenas um exame e passa a ser uma trajetória. Essa diferença é maior do que parece. Em uma imagem estática, o sistema procura um padrão presente. Em um fluxo longitudinal, precisa distinguir variação normal de mudança clinicamente relevante, considerar o que vinha acontecendo antes e decidir quando uma alteração merece chamar a atenção de alguém. A unidade de análise já não é somente o dado; é o dado no tempo.

É importante, entretanto, não chamar todo dispositivo conectado de inteligência artificial. Um sensor mede; um software transmite; uma regra pode comparar o valor com um limite; um modelo de aprendizado de máquina pode estimar risco a partir de múltiplos sinais. Essas camadas podem coexistir, mas não são equivalentes. Um relógio capaz de registrar frequência cardíaca não se torna "inteligente" apenas por produzir muitos dados. O componente de IA aparece quando há inferência algorítmica capaz de classificar, predizer, priorizar ou adaptar uma saída a partir desses dados. Manter essa distinção evita que inovação de infraestrutura seja confundida com evidência algorítmica.

Também evita outra simplificação: mais dados não significam necessariamente mais conhecimento. Séries contínuas contêm ruído, perdas de conexão, períodos em que o dispositivo não foi usado, mudanças de posicionamento, artefatos de movimento e diferenças entre equipamentos. O silêncio do sensor pode significar estabilidade, bateria descarregada, abandono do dispositivo ou dificuldade de acesso à internet. Se o modelo trata ausência como normalidade, um problema técnico pode se transformar em falsa tranquilidade. Se trata qualquer desvio como risco, o sistema pode produzir uma quantidade de alertas impossível de absorver.

Essa é uma das razões pelas quais o campo ainda apresenta distância importante entre demonstração técnica e utilidade clínica. A revisão sistemática de Wolber et al. (2026), dedicada a estudos prospectivos de monitoramento remoto de condições crônicas com aprendizado de máquina, encontrou 76 estudos elegíveis entre 6.668 registros. Embora as aplicações incluíssem doenças neurológicas, diabetes, doença pulmonar e condições cardíacas, 73,7% dos estudos foram classificados com alto risco de viés, principalmente por problemas analíticos. O resultado não diminui o potencial do monitoramento; mostra que um fluxo contínuo de dados pode parecer tecnologicamente maduro antes de sua evidência clínica amadurecer na mesma velocidade.

Em séries temporais, a validação exige cuidados adicionais. Dividir aleatoriamente milhares de observações do mesmo paciente entre treino e teste pode produzir a impressão de generalização quando o modelo, na prática, está reconhecendo características daquela mesma pessoa. Um sistema destinado a acompanhar pacientes novos precisa ser avaliado em indivíduos que não participaram do desenvolvimento e, idealmente, em contextos diferentes. Além disso, o horizonte temporal precisa corresponder à ação pretendida. Detectar uma piora quinze minutos antes de ela se tornar óbvia pode ter valor em determinada emergência; antecipá-la por três dias pode ser decisivo em uma doença crônica tratável em casa.

A fibrilação atrial oferece um exemplo particularmente útil porque mostra a evolução da evidência do monitoramento em larga escala para a avaliação randomizada. No Apple Heart Study, Perez et al. (2019) acompanharam 419.297 participantes por meio de um algoritmo de notificação de pulso irregular. Apenas 0,52% receberam uma notificação durante mediana de 117 dias. Entre os 450 participantes que devolveram um patch de eletrocardiograma com dados analisáveis depois do alerta, fibrilação atrial foi identificada em 34%. Quando uma nova

notificação ocorria simultaneamente ao registro eletrocardiográfico, o valor preditivo positivo para fibrilação atrial foi de 0,84.

O estudo mostrou que um dispositivo de consumo podia participar de uma estratégia de detecção em escala antes difícil de imaginar, mas também expôs limites importantes. Era um estudo de braço único, dependia de pessoas que possuíam dispositivos compatíveis e falavam inglês e avaliava principalmente o que acontecia depois de uma notificação. Encontrar mais episódios não responde, por si só, se rastrear aquela população melhora desfechos. A pergunta clinicamente mais difícil começa depois da detecção: quais episódios justificam investigação, anticoagulação ou outra intervenção e qual é o balanço entre prevenção de eventos e danos produzidos por sobrediagnóstico ou tratamento?

Em 2026, o ensaio EQUAL aproximou essa questão de um desenho experimental mais robusto. Van Steijn et al. randomizaram 437 pessoas com 65 anos ou mais e risco elevado de acidente vascular cerebral para seis meses de monitoramento com smartwatch ou cuidado habitual. Nova fibrilação atrial foi diagnosticada em 9,6% no grupo monitorado e em 2,3% no controle, diferença absoluta de 7,3 pontos percentuais. Os eletrocardiogramas de uma derivação eram revisados por uma equipe remota em até 24 horas, mostrando que o benefício diagnóstico dependia não apenas do relógio, mas de uma cadeia organizada de confirmação e resposta (VAN STEIJN et al., 2026).

O ensaio é importante justamente pelo que demonstra e pelo que ainda não demonstra. Ele sustenta que o monitoramento com smartwatch pode aumentar e acelerar a detecção de fibrilação atrial em uma população selecionada de alto risco. Não estabelece, sozinho, que essa estratégia reduza acidentes vasculares cerebrais, mortalidade ou incapacidade. Essa diferença é central para o cuidado contínuo: quanto mais fácil se torna observar um fenômeno, maior a tentação de transformar a capacidade de encontrá-lo em argumento para intervir. A tecnologia pode ampliar a janela de observação sem resolver automaticamente a incerteza sobre o que deve ser feito.

A própria noção de "contínuo" merece cautela. Muitos dispositivos não medem todos os fenômenos o tempo inteiro; amostram sinais em determinados intervalos, dependem de contato adequado com a pele ou exigem que o usuário ative uma função. Outros produzem uma quantidade de dados tão grande que a análise só é possível depois de compressão, filtragem e seleção de características. O que chega ao

profissional já é uma interpretação de uma sequência de escolhas técnicas. Entre o corpo e o alerta existem sensor, algoritmo, transmissão, armazenamento, modelo e interface.

Por isso, um biomarcador digital não deveria ser tratado como se fosse automaticamente equivalente a um biomarcador laboratorial consolidado. Número de passos, variabilidade da frequência cardíaca, padrão de sono, temperatura periférica e aceleração podem carregar informação útil, mas seu significado depende de contexto, dispositivo, comportamento e finalidade. Uma queda de atividade pode anteceder piora clínica; pode também refletir chuva, viagem, dor musculoesquelética ou o fato de o relógio ter ficado sobre a mesa. A inteligência do sistema está menos em medir mais e mais em conseguir separar mudança relevante de variação cotidiana sem ocultar incerteza.

Quando acompanhar significa intervir

Monitorar só se transforma em cuidado quando existe um circuito de resposta. Essa afirmação parece óbvia, mas é uma das fronteiras mais difíceis da saúde digital. Um algoritmo pode antecipar risco com boa discriminação e ainda produzir pouco valor se o aviso chegar a uma equipe sem tempo, protocolo ou autoridade para agir. O inverso também ocorre: uma informação relativamente simples pode ser valiosa quando chega cedo, é compreensível e desencadeia uma intervenção eficaz. O sistema clínico, portanto, não começa no sensor e termina no escore. Ele termina na consequência produzida pela informação.

A insuficiência cardíaca tornou-se um laboratório importante para essa discussão. Peso, frequência cardíaca, frequência respiratória, atividade, sono e outros sinais podem se modificar antes de uma descompensação evidente. Modelos multivariados tentam reconhecer essas mudanças antes que o paciente apresente dispneia intensa ou precise ser hospitalizado. A promessa é criar uma janela para ajustar diuréticos, revisar adesão, antecipar consulta ou investigar outra causa de deterioração. A dificuldade é que cada alerta abre uma sequência de decisões que consome atenção e pode produzir tanto benefício quanto intervenção desnecessária.

O programa LINK-HF2 foi construído para testar precisamente essa passagem. Na fase piloto, publicada no Journal of the American Medical Informatics Association, profissionais responderam a 95% das notificações em até 24 horas, mas apenas 26,7% resultaram em ação clínica. Entrevistas mostraram que responder aos avisos exigia revisar prontuário, contatar o paciente, articular-se com outros

profissionais e julgar se o sinal merecia confiança. Em outras palavras, o algoritmo prometia reduzir incerteza fisiológica, mas criava trabalho cognitivo e organizacional que precisava ser incorporado ao desenho da intervenção (SIDERIS et al., 2024).

Os resultados randomizados apresentados em 2026 no suplemento do JACC tornam a história ainda mais instrutiva. Em cinco centros do sistema de saúde dos veteranos dos Estados Unidos, pacientes recentemente internados por insuficiência cardíaca utilizaram um sensor por até 90 dias. No grupo de intervenção, notificações geradas por análise preditiva eram enviadas à equipe; no controle, os sinais eram coletados, mas as notificações não eram comunicadas. Entre 171 participantes analisados, os alertas levaram a maior vigilância, ajustes de medicamentos, consultas antecipadas e alguns encaminhamentos para emergência (SIDERIS et al., 2026).

Apesar dessa atividade assistencial, a hospitalização por insuficiência cardíaca foi semelhante entre os grupos: razão de riscos de 0,97, com intervalo de confiança amplo e valor de p de 0,92. Houve redução na ocorrência cumulativa de novas notificações de piora fisiológica, interpretada pelos autores como maior estabilidade, mas o estudo não demonstrou benefício no desfecho clínico primário. Como os resultados foram publicados em formato de suplemento de congresso e a amostra foi limitada, a conclusão apropriada é modesta: agir sobre sinais preditivos pareceu factível e modificou condutas, mas ainda não ficou demonstrado que esse circuito reduza hospitalizações.

Esse tipo de resultado é valioso porque impede uma narrativa excessivamente linear. Detectar antes não garante tratar melhor; tratar mais cedo não garante modificar o desfecho; e modificar um marcador intermediário não garante melhorar a experiência ou a sobrevida do paciente. Entre cada uma dessas etapas existem hipóteses que precisam ser testadas. Em algumas doenças, a janela antecipada pode permitir uma intervenção altamente efetiva. Em outras, o sinal pode aparecer quando a trajetória já está determinada ou pode não indicar qual intervenção é a correta.

Também não se deve atribuir à inteligência artificial benefícios que pertencem ao programa de cuidado como um todo. Monitoramento remoto pode funcionar com regras simples, questionários, telefonemas, teleconsultas e protocolos conduzidos por enfermeiros ou médicos. O ensaio de Yende et al. (2026) torna essa observação especialmente importante. Em 1.286 pessoas após internação por sepse ou infecção respiratória, quatro estratégias de monitoramento remoto, combinando intensidades

de questionários e diferentes equipes de resposta, não aumentaram os dias vivos em casa em 90 dias em comparação ao cuidado habitual.

O achado mais desconfortável apareceu entre participantes com 65 anos ou mais, nos quais determinadas estratégias de monitoramento foram associadas a menos dias em casa e mais reinternações. O estudo não era um ensaio de IA para interpretação fisiológica, e justamente por isso funciona como controle conceitual: digitalizar o acompanhamento e aumentar a frequência de contato não são intervenções intrinsecamente benéficas. Monitoramento pode produzir avaliações adicionais, deslocamentos, ansiedade e hospitalizações que talvez não ocorressem sob outra estratégia. O efeito depende da população, do sinal observado e de como o sistema reage (YENDE et al., 2026).

Esse problema pode ser descrito como limiar de ação. Se o sistema é muito sensível, encontra alterações precoces e inespecíficas e aumenta a quantidade de pessoas investigadas. Se é muito conservador, reduz falsos alertas, mas pode perder a oportunidade de intervir. Não existe um limiar universalmente correto porque o custo do erro muda conforme a situação. Uma arritmia potencialmente grave, uma oscilação de peso, uma queda de saturação e uma redução de atividade não possuem a mesma urgência nem a mesma relação entre falso positivo e falso negativo.

O cuidado contínuo ainda introduz uma forma particular de fadiga. Em sistemas pontuais, um profissional recebe uma recomendação durante uma consulta. Em monitoramento longitudinal, a mesma pessoa pode gerar dezenas ou centenas de eventos ao longo de semanas. A pergunta deixa de ser somente se cada alerta é "correto" e passa a incluir quantos alertas uma equipe consegue revisar sem perder capacidade de julgamento. Métricas como sensibilidade precisam ser acompanhadas de carga de alertas, tempo de resposta, taxa de ações, capacidade de resolução e consequência das intervenções desencadeadas.

Há também um efeito sobre o paciente. Ser acompanhado continuamente pode transmitir segurança, mas pode transformar o corpo em fonte permanente de notificações. Algumas pessoas se beneficiam ao compreender tendências; outras podem desenvolver preocupação excessiva diante de oscilações sem significado. O desenho responsável precisa definir quais sinais são mostrados diretamente ao usuário, quais devem passar por validação clínica e como comunicar incerteza. Um alerta de "possível alteração" não deveria ser apresentado com a mesma autoridade de um diagnóstico confirmado.

A adesão, nesse contexto, deixa de ser um detalhe operacional e passa a compor a validade do sistema. Modelos desenvolvidos com dados de pessoas que usam o dispositivo de forma consistente podem funcionar pior justamente entre aquelas que têm maior dificuldade de manter carregamento, conectividade ou rotina de uso. Se a perda de dados for sistemática, o modelo pode tornar-se mais confiável para os usuários mais organizados, saudáveis ou tecnologicamente favorecidos e menos útil para quem enfrenta maior complexidade social ou clínica.

A revisão sistemática de Wu et al. (2026), disponibilizada em acesso antecipado em agosto de 2026, mostra que a exclusão digital entre adultos mais velhos não se reduz a "não saber usar tecnologia". Ela emerge da combinação de desvantagens socioeconômicas, vulnerabilidades de saúde e barreiras contextuais e se associa a piores resultados de saúde e qualidade de vida. A implicação para monitoramento inteligente é direta: se receber cuidado passa a depender de smartphone compatível, conexão estável, visão, destreza e letramento digital, a própria arquitetura do programa pode selecionar quem consegue ser acompanhado.

Por isso, oferecer alternativa não digital não é uma concessão contrária à inovação. É parte da segurança e da equidade do desenho. Programas maduros precisam prever treinamento, suporte técnico, substituição de equipamentos, canais humanos e formas de reconhecer rapidamente quando a ausência de dados é falha do sistema e não estabilidade clínica. A tecnologia pode levar cuidado para fora do hospital, mas não deve levar para o paciente toda a responsabilidade por manter a infraestrutura funcionando.

Quando o cuidado contínuo ganha escala

A mesma lógica que acompanha um indivíduo pode ser ampliada para milhares de pessoas. Em saúde populacional, modelos percorrem registros eletrônicos, dados administrativos e outras fontes para estimar quem apresenta maior probabilidade de internação, complicação, abandono de tratamento ou necessidade de acompanhamento intensivo. A promessa é usar recursos escassos de forma mais antecipatória: em vez de esperar que o paciente retorne em crise, o sistema identifica grupos que merecem busca ativa, revisão de tratamento ou coordenação adicional.

Essa mudança de escala altera a natureza da decisão. Um escore individual responde "qual é o risco desta pessoa?". Um sistema populacional também responde, implicitamente, "quem será visto primeiro?". Quando uma equipe dispõe de cem vagas e o algoritmo ordena mil pacientes, a previsão deixa de ser apenas informação e

passa a participar da distribuição de atenção. Por isso, escolher o desfecho que o modelo prevê torna-se uma decisão organizacional com consequências clínicas.

O estudo de Raldow et al. (2024) mostra como a IA pode adquirir valor quando conectada a uma intervenção populacional concreta. Em 48 clínicas de atenção primária da UCLA Health, um modelo desenvolvido pelo próprio sistema identificava pacientes em faixas elevadas de risco. A implementação foi realizada em desenho randomizado por conglomerados do tipo stepped-wedge e associada a um programa proativo de gestão do cuidado. Na análise agrupada de 3.007 pacientes de alto e altíssimo risco, houve redução de nível de 27% nas internações potencialmente evitáveis após a intervenção.

O resultado precisa ser lido com precisão. As visitas potencialmente evitáveis ao pronto atendimento não apresentaram redução equivalente, e as quedas em todas as hospitalizações e visitas de emergência não alcançaram significância estatística convencional. Além disso, o mecanismo não foi "o algoritmo evitar a internação". O modelo selecionou pessoas; enfermeiros, profissionais de serviço social e equipes clínicas revisaram necessidades, organizaram planos e atuaram sobre barreiras. O efeito, quando apareceu, emergiu da combinação entre estratificação e capacidade humana de oferecer cuidado adicional (RALDOW et al., 2024).

Esse exemplo ajuda a estabelecer uma regra para saúde populacional: desempenho preditivo só é parte do valor. Um modelo pode identificar corretamente pessoas de alto risco e ainda fracassar se a rede não tiver consultas, transporte, assistência farmacêutica, atenção domiciliar ou profissionais para responder. A IA pode tornar uma fila mais sofisticada sem aumentar a capacidade do sistema. Em cenários de escassez, isso talvez apenas revele com maior precisão um problema que continua sem solução.

A escolha do alvo também importa. Obermeyer et al. (2019) demonstraram esse problema ao estudar um algoritmo comercial usado para selecionar pacientes que receberiam programas adicionais de cuidado. O sistema utilizava gasto em saúde como aproximação de necessidade clínica. Como pacientes negros historicamente recebiam menos recursos para níveis semelhantes de doença, o algoritmo subestimava suas necessidades. A correção do viés aumentaria de 17,7% para 46,5% a proporção de pacientes negros entre aqueles selecionados para ajuda adicional.

O episódio é frequentemente citado como exemplo de viés algorítmico, mas também ensina algo específico sobre monitoramento populacional. Quando a IA

organiza cuidado, uma variável aparentemente objetiva pode representar outra coisa. Custo não é doença; número de consultas não é necessidade; ausência de retorno não é ausência de problema. Modelos populacionais precisam ser avaliados não apenas por sua capacidade de prever o alvo escolhido, mas pela legitimidade desse alvo como aproximação do que o sistema pretende proteger.

No Brasil, essa discussão encontra um sistema com potencial e limitações próprias. A revisão de escopo de Botero et al. (2025) identificou 27 estudos sobre aplicações de inteligência artificial na Atenção Primária à Saúde brasileira entre 981 registros recuperados. Aprendizado de máquina e deep learning foram utilizados para estimativas e previsões, rastreamento e diagnóstico de condições e análise de determinantes relacionados à cobertura em saúde. Entre os benefícios relatados estavam otimização de recursos, diagnóstico precoce, estratificação de risco e facilitação do acesso; entre os desafios, disponibilidade de dados e acesso insuficiente ao tratamento.

Esses dois últimos desafios são inseparáveis. Um modelo treinado com dados incompletos pode produzir estratificação imprecisa; um modelo preciso inserido em uma rede sem capacidade de atender pode produzir pouca mudança assistencial. Na atenção primária, o valor de antecipar risco depende de longitudinalidade, território e coordenação. A equipe precisa conhecer o que aconteceu depois do alerta, atualizar o plano e manter vínculo. Sem esse retorno, a inteligência artificial corre o risco de operar como mecanismo sofisticado de triagem sem completar o ciclo do cuidado.

Interoperabilidade torna-se especialmente importante quando o acompanhamento atravessa níveis assistenciais. Dados de um wearable podem indicar deterioração, mas a equipe de atenção primária precisa saber que o paciente foi à emergência; o hospital precisa conhecer medicamentos e histórico; o serviço remoto precisa registrar a intervenção de modo que outros profissionais a encontrem. Se cada parte enxerga apenas seu próprio banco, a tecnologia aumenta a quantidade de dados sem construir uma história clínica comum.

O tempo também muda o problema da governança. Um modelo diagnóstico pode ser avaliado em uma versão definida. Um sistema de acompanhamento permanece exposto a meses ou anos de mudança: pacientes envelhecem, tratamentos modificam a fisiologia, dispositivos recebem atualizações, protocolos são alterados e padrões epidemiológicos se deslocam. Monitorar o paciente exige também monitorar o monitor. Taxas de alerta, distribuição dos sinais, perdas de dados, desempenho por

subgrupos e consequências das ações precisam ser acompanhadas depois da implantação.

Esse princípio ajuda a definir o que seria um sistema realmente inteligente de cuidado contínuo. Não é o sistema que mais coleta, nem aquele que produz o maior número de previsões. É aquele capaz de selecionar informação clinicamente relevante, reconhecer quando sua própria confiança é insuficiente, encaminhar o sinal para uma resposta possível e aprender institucionalmente com os resultados sem modificar-se de forma descontrolada. A inteligência relevante para a saúde não está apenas no modelo; está na arquitetura que liga observação, interpretação, ação e revisão.

Há, portanto, uma mudança conceitual importante ao sair da decisão pontual para o acompanhamento longitudinal. No capítulo anterior, a pergunta central era como uma previsão deveria influenciar uma escolha. Aqui, a dificuldade está em decidir quando uma mudança ao longo do tempo merece tornar-se uma decisão. Essa fronteira é menos visível, mas pode ser ainda mais poderosa: quem define o alerta define, em parte, o momento em que um paciente entra novamente no campo de atenção do sistema.

O monitoramento contínuo também prepara a próxima transformação. Até aqui, grande parte dos exemplos depende de sinais relativamente estruturados: ritmo, frequência, atividade, valores laboratoriais, eventos no prontuário e escores de risco. O cuidado real, porém, produz uma camada adicional de informação em linguagem: evoluções, mensagens, relatos de sintomas, conversas, orientações e documentos. Modelos generativos e multimodais prometem combinar essas fontes e transformar grandes volumes de dados longitudinais em sínteses, respostas e recomendações.

Essa promessa muda novamente a interface entre inteligência artificial e cuidado. Em vez de apenas vigiar um sinal e emitir um alerta, o sistema pode passar a explicar o que mudou, resumir semanas de evolução, conversar com pacientes e profissionais e produzir conteúdo novo a partir de informações heterogêneas. A passagem do monitoramento para a linguagem amplia a utilidade potencial, mas também cria um problema diferente: quando o sistema deixa de apenas calcular e começa a formular narrativas clínicas, como distinguir síntese confiável de informação plausível, porém inventada? A dificuldade deixa de estar somente no sinal captado e passa a incluir a fidelidade da própria narrativa produzida sobre o paciente.

CAPÍTULO 5
QUANDO A IA PASSA A PRODUZIR: MODELOS GENERATIVOS, MULTIMODAIS E LINGUAGEM CLÍNICA

Quando a linguagem vira interface

O capítulo anterior terminou diante de uma mudança de natureza. Até ali, grande parte dos sistemas acompanhava sinais, atualizava riscos ou ordenava pacientes a partir de informações relativamente estruturadas. Modelos generativos introduzem outra possibilidade: eles recebem linguagem e devolvem linguagem. Podem resumir um prontuário, redigir uma evolução, responder a uma pergunta, organizar hipóteses, transformar uma conversa em nota clínica ou combinar texto com imagens e outros dados. A inteligência artificial deixa de aparecer apenas como número, classe ou alerta e passa a participar da forma como o caso é narrado.

Essa mudança de interface é mais profunda do que parece. Um escore de 0,82 é claramente uma saída computacional; uma resposta bem escrita pode ser recebida como se fosse uma explicação. A linguagem carrega tom, encadeamento, justificativa e aparência de intenção. Por isso, modelos de linguagem de grande escala não precisam estar certos para parecer convincentes. Eles produzem sequências linguisticamente plausíveis a partir de padrões aprendidos em grandes conjuntos de dados; não consultam, por definição, uma base de fatos a cada frase nem possuem um mecanismo interno equivalente ao dever clínico de verificar a verdade antes de afirmá-la. Essa diferença entre fluência e factualidade é uma das chaves para compreender seu uso em saúde.

Os primeiros resultados de grande visibilidade ajudaram a demonstrar por que esses modelos despertaram tanto interesse. Singhal et al. (2023), ao desenvolverem e avaliarem o Med-PaLM, mostraram que a ampliação de escala e o ajuste para o domínio médico melhoravam o desempenho em conjuntos de perguntas clínicas. O estudo também foi cuidadoso em um ponto que frequentemente se perde na divulgação: desempenho em perguntas de prova não bastava. Na avaliação de respostas a perguntas médicas de consumidores, os próprios autores observaram que modelos fortes ainda podiam produzir conteúdo inadequado para um domínio em que erro e segurança importam. A competência linguística havia avançado; a distância entre responder e cuidar permanecia aberta.

Moor et al. (2023) propuseram chamar de inteligência artificial médica generalista a possibilidade de modelos fundacionais que não fossem treinados apenas para uma tarefa estreita, mas capazes de combinar diferentes modalidades e produzir saídas diversas. A ideia é importante porque rompe com a arquitetura clássica de um algoritmo para cada problema. Um mesmo sistema poderia, em princípio, resumir dados, interpretar uma imagem em conjunto com a história, responder perguntas e apoiar diferentes etapas do cuidado. Mas o conceito descreve um horizonte tecnológico, não uma certificação de competência geral. Quanto maior a flexibilidade de uma ferramenta, maior também o número de situações em que ela pode ser utilizada fora do contexto em que sua capacidade foi realmente demonstrada.

Essa distinção ficou especialmente clara quando os modelos começaram a ser testados com médicos, e não apenas contra bancos de questões. Goh et al. (2024) randomizaram 50 médicos para resolver casos clínicos complexos com acesso a um grande modelo de linguagem mais recursos convencionais ou apenas com recursos convencionais. A pontuação mediana de raciocínio diagnóstico foi de 76% no grupo com o modelo e 74% no controle, diferença ajustada de apenas 2 pontos percentuais e sem significância estatística. Curiosamente, o modelo avaliado sozinho obteve desempenho superior ao grupo controle. O resultado expõe um problema que benchmarks isolados não mostram: uma ferramenta pode ter informação útil e, ainda assim, essa informação não se transformar automaticamente em melhor raciocínio quando entra na interação com uma pessoa.

Um ensaio posterior do mesmo grupo encontrou um quadro diferente. Goh et al. (2025) estudaram 92 médicos em tarefas de raciocínio de manejo, nas quais informações eram apresentadas progressivamente e os participantes precisavam decidir exames, tratamentos e próximos passos. O grupo com GPT-4 e recursos convencionais alcançou pontuação média 6,5 pontos percentuais maior que o grupo com recursos convencionais, embora tenha gasto, em média, cerca de dois minutos adicionais por caso. A evidência, portanto, não é simplesmente contraditória. Diagnóstico diferencial e manejo são tarefas diferentes; a forma como a informação é apresentada e como o modelo é utilizado também muda. O efeito da IA precisa ser descrito para a tarefa em que foi observado.

O estudo de Qazi et al. (2026) acrescentou uma variável ainda mais interessante: treinamento do usuário. Sessenta médicos licenciados de diferentes instituições do Paquistão realizaram previamente um programa de 20 horas sobre capacidades, limitações, estratégias de uso e avaliação crítica de modelos de linguagem. Entre os 58

que concluíram o ensaio, aqueles que receberam acesso ao GPT-4o além dos recursos convencionais atingiram pontuação média de raciocínio diagnóstico de 71,4%, contra 42,6% no grupo que utilizou apenas recursos convencionais, diferença ajustada de 27,5 pontos percentuais. O tempo por caso foi semelhante. Ainda assim, em análise exploratória, o modelo sozinho superou, em média, a combinação médico-modelo. O dado mais útil não é escolher um vencedor; é perceber que alfabetização em IA pode modificar substancialmente a interação.

Os três ensaios, lidos em conjunto, desmontam duas simplificações. A primeira é imaginar que um modelo que se sai bem sozinho necessariamente melhorará o profissional. A segunda é concluir, a partir de um estudo negativo, que a colaboração humano-máquina não oferece valor. O desempenho conjunto depende de treinamento, formulação da tarefa, experiência, tempo, confiança e capacidade de incorporar ou rejeitar a resposta. A unidade real de avaliação não é apenas o modelo nem apenas o médico, mas a interação entre ambos dentro de um processo específico.

Hager et al. (2024) reforçaram esse ponto ao avaliar grandes modelos em tarefas que se aproximavam mais da dinâmica clínica. Quando precisavam lidar com informação incompleta, decidir quais dados solicitar, seguir diretrizes e manter consistência diante de pequenas mudanças de instrução, os modelos apresentavam limitações que não apareciam com a mesma clareza em provas fechadas. A medicina não entrega ao sistema uma pergunta perfeitamente formulada com todas as variáveis necessárias. Muitas vezes, parte do trabalho é descobrir o que ainda precisa ser perguntado.

Isso coloca a noção de prompt em seu devido lugar. Formular uma instrução clara pode melhorar uma resposta, assim como fazer uma boa pergunta melhora uma consulta a qualquer fonte. Mas segurança clínica não pode depender da existência de uma frase secreta que faça o modelo funcionar corretamente. Se uma ferramenta exige habilidade extraordinária para não produzir erro perigoso, o problema é também de projeto e de implementação. Treinamento do usuário importa, mas não substitui validação, desenho de interface, limites de uso e mecanismos de revisão.

O mesmo vale para sistemas conectados a mecanismos de busca ou recuperação de documentos. Dar ao modelo acesso a diretrizes, protocolos locais ou trechos do prontuário pode reduzir a dependência exclusiva do conteúdo aprendido durante o treinamento e permitir respostas ancoradas em fontes identificáveis. Isso melhora uma parte do problema, mas não transforma geração em cópia fiel. O sistema ainda

seleciona, combina e formula. Pode recuperar o documento correto e interpretá-lo mal, omitir uma exceção ou construir uma conclusão mais forte do que a evidência fornecida. Em saúde, acesso à fonte e fidelidade à fonte são requisitos diferentes.

A linguagem, portanto, é ao mesmo tempo vantagem e risco. Ela permite que profissionais interajam com sistemas complexos sem aprender uma interface técnica para cada função. Permite também que pacientes formulem dúvidas em suas próprias palavras. Mas a mesma naturalidade pode apagar sinais de incerteza. Quanto mais a ferramenta se parece com um interlocutor competente, maior a necessidade de lembrar que sua autoridade precisa vir da evidência, não do estilo da resposta.

Quando escrever também pode mudar o cuidado

Os usos mais próximos da rotina talvez não sejam aqueles em que o modelo tenta resolver um diagnóstico difícil, mas os que trabalham sobre a enorme quantidade de texto que já existe na assistência. Prontuários contêm evoluções, laudos, prescrições, mensagens, antecedentes, resultados e cópias de informações produzidas em momentos diferentes. Ler tudo isso demanda tempo; escrever tudo novamente também. A capacidade de resumir e redigir tornou a documentação um dos primeiros territórios em que a IA generativa começou a ser incorporada a sistemas clínicos reais.

Resumir, contudo, não é uma atividade neutra. Toda síntese decide o que permanece e o que desaparece. Em um romance, uma omissão pode empobrecer a história; em um prontuário, pode retirar a informação que explicava uma alergia, uma contraindicação ou a razão pela qual determinado tratamento foi suspenso. Também é possível errar temporalidade, atribuir um achado antigo ao episódio atual ou transformar hipótese em fato. A qualidade de uma síntese clínica não depende apenas de ser curta e legível, mas de preservar relações que mudam decisões.

Van Veen et al. (2024) avaliaram oito modelos adaptados em quatro tarefas de sumarização clínica: laudos radiológicos, perguntas de pacientes, notas de evolução e diálogos entre médico e paciente. Em um estudo de leitura com dez médicos, as melhores sínteses produzidas pelos modelos foram consideradas equivalentes às sínteses de especialistas em 45% das avaliações e superiores em 36%. O resultado demonstra capacidade relevante, mas o próprio estudo realizou uma análise de segurança que identificou erros e informações fabricadas tanto em saídas de modelos quanto em textos humanos. A conclusão responsável não é que resumir deixou de exigir revisão; é que a geração automática se tornou suficientemente competente para merecer avaliação dentro do fluxo de trabalho.

Esse ponto aparece com ainda mais clareza nas mensagens enviadas por pacientes. Tai-Seale et al. (2024) estudaram a introdução de rascunhos gerados por IA dentro do prontuário eletrônico para respostas a mensagens. A hipótese intuitiva era de economia de tempo. O resultado foi mais complexo: o acesso aos rascunhos se associou a maior tempo de leitura das mensagens, não produziu redução significativa do tempo de resposta e levou a respostas mais longas. Os médicos, por outro lado, relataram valor em receber um ponto de partida que percebiam como mais compassivo. A tecnologia modificou a escrita sem entregar automaticamente o ganho de produtividade que se poderia esperar.

Essa diferença é importante porque eficiência possui várias dimensões. Um rascunho pode não reduzir segundos medidos na tela e ainda diminuir a dificuldade de iniciar uma resposta; pode aumentar o tempo de revisão porque o profissional precisa conferir cada afirmação; pode produzir mensagens mais claras, mas excessivamente longas. Avaliar apenas velocidade perderia parte do efeito. Avaliar apenas satisfação também. Ferramentas generativas alteram esforço cognitivo, atenção, linguagem e distribuição do trabalho, e esses efeitos nem sempre caminham na mesma direção.

Os chamados escribas ambientais levam essa proposta um passo adiante. Em vez de o profissional ditar uma nota depois da consulta, o sistema capta o áudio do encontro, identifica falas relevantes e gera um rascunho estruturado. A promessa é devolver atenção à conversa ao reduzir a necessidade de olhar continuamente para a tela. Mas esse processo não é simples transcrição. Para transformar diálogo em nota, o modelo seleciona conteúdo, organiza seções, resolve referências, interpreta abreviações e, em alguma medida, decide o que considera clinicamente saliente. Quanto mais o produto final parece pronto, maior pode ser a tentação de revisar menos.

O primeiro ensaio randomizado pragmático de grande porte com duas plataformas de escriba ambiental, publicado por Lukac et al. (2025), ajuda a colocar a promessa em perspectiva. Duzentos e trinta e oito médicos ambulatoriais de 14 especialidades foram distribuídos entre DAX Copilot, Nabla ou cuidado usual. Na análise por intenção de tratar, usuários do Nabla apresentaram redução de 9,5% no tempo de escrita por nota em comparação ao controle; no grupo DAX, a redução de 1,7% não foi estatisticamente significativa. Medidas secundárias sugeriram melhora em aspectos de carga de trabalho e exaustão para as duas plataformas, mas os autores destacaram que esses achados precisavam de confirmação.

O ensaio também mostrou por que uma categoria de produto não deve ser tratada como se fosse uma tecnologia homogênea. Duas ferramentas com finalidade semelhante produziram resultados diferentes no desfecho primário. A utilização real ficou abaixo de metade das consultas em ambos os grupos. E os médicos relataram inexatidões clinicamente significativas de forma ocasional, incluindo omissões, problemas estruturais e erros de resolução de pronomes. Não basta perguntar se "escribas de IA funcionam". É necessário saber qual sistema, em qual especialidade, com qual taxa de uso, em que tipo de encontro e com qual procedimento de revisão.

Há um efeito de segunda ordem que merece atenção: texto gerado pode se tornar dado clínico. Uma nota aprovada entra no prontuário e poderá ser lida por outro profissional, utilizada em auditoria, alimentar um resumo futuro ou servir como entrada para outro modelo. Se uma informação inventada passa despercebida e adquire aparência de registro oficial, o erro deixa de existir apenas por alguns segundos na tela. Ele ganha persistência. Em sistemas interconectados, geração e documentação formam uma cadeia na qual uma pequena distorção pode ser repetida e reforçada.

Por isso, a revisão humana precisa ser desenhada para encontrar erro, e não apenas formalmente prevista. Ler um rascunho longo depois de uma consulta intensa é diferente de escrever a própria nota. O texto gerado pode criar ancoragem: quando a frase está pronta e soa correta, é mais difícil perceber o dado que faltou. Interfaces podem ajudar destacando informações extraídas diretamente do encontro, sinalizando trechos incertos ou permitindo rastrear cada afirmação até a fonte correspondente. A melhor supervisão não é aquela que simplesmente transfere ao profissional toda a responsabilidade depois que o sistema escreveu; é a que torna a verificação cognitivamente possível.

O mesmo princípio se aplica à comunicação com pacientes. Um modelo pode ajudar a traduzir termos técnicos, organizar orientações ou adaptar uma mensagem para linguagem mais acessível. Isso pode ser valioso, sobretudo quando o profissional mantém controle do conteúdo. Mas comunicação clínica não é apenas transmitir informação correta. É reconhecer urgência, saber o que não pode ser resolvido por mensagem, considerar contexto, acolher dúvida sem criar falsa tranquilidade e indicar quando uma recomendação depende de avaliação presencial. A qualidade linguística precisa permanecer subordinada à segurança do percurso assistencial.

A revisão sistemática de Busch et al. (2025), já discutida anteriormente, encontrou justamente essa mistura de oportunidades e limitações em aplicações de

grandes modelos de linguagem no cuidado ao paciente. Resposta a perguntas, produção e adaptação de informação, sumarização, tradução e documentação aparecem entre os usos frequentes, ao lado de problemas de incorreção, incompletude, viés, transparência e reprodutibilidade. A IA generativa se aproxima rapidamente do trabalho cotidiano porque escreve bem; essa proximidade torna ainda mais importante medir o que acontece depois que o texto é aceito e usado.

Quando texto, imagem e contexto se encontram

A assistência, porém, nunca foi puramente textual. Um caso pode reunir radiografia, anatomia patológica, eletrocardiograma, exames laboratoriais, medicações, histórico, dados genéticos e relato do paciente. Durante décadas, boa parte da inteligência artificial biomédica tratou essas fontes separadamente: um modelo para imagem, outro para texto, outro para risco. A multimodalidade procura aproximar o sistema computacional da forma como o raciocínio clínico combina evidências diferentes.

Acosta et al. (2022) já descreviam a inteligência artificial biomédica multimodal como uma resposta à expansão simultânea de registros eletrônicos, imagens, sensores, biobancos e dados ômicos. O ganho potencial não está em acumular modalidades por si só, mas em explorar relações entre elas. Uma imagem pode adquirir significado diferente quando lida com idade, sintomas e exames prévios; uma alteração laboratorial pode importar mais quando coincide com uma tendência de sinais vitais. O desafio é que integração também cria novos pontos de falha. Dados podem estar ausentes, desalinhados no tempo ou discordar entre si.

Os modelos visão-linguagem tornaram essa integração mais visível. Zhang et al. (2024) apresentaram o BiomedGPT, um modelo fundacional aberto capaz de trabalhar com diferentes tarefas biomédicas de imagem e linguagem. Em avaliações humanas de radiologia, o sistema demonstrou desempenho competitivo em resposta a perguntas, geração de laudos e sumarização, com taxas de erro que variaram conforme a tarefa. O estudo mostra que uma única arquitetura pode adquirir versatilidade que antes exigia modelos separados. Não demonstra, contudo, que um modelo generalista possa assumir autonomamente todas essas funções na clínica.

A patologia oferece um exemplo ainda mais concreto. Lu et al. (2024) desenvolveram o PathChat, combinando um codificador visual especializado em histopatologia com um grande modelo de linguagem e treinamento sobre centenas de milhares de instruções envolvendo imagem e texto. Em questões diagnósticas e

avaliações abertas por especialistas, o sistema apresentou desempenho superior a comparadores multimodais em diversos cenários. Os próprios autores o descrevem como prova de conceito para educação, pesquisa e apoio humano no processo decisório. Essa qualificação importa: responder adequadamente a imagens selecionadas em um estudo não equivale a assumir responsabilidade por um laudo em serviço real.

A diferença entre demonstração e assistência clínica também aparece em modelos generalistas baseados em registros. Liu et al. (2025) apresentaram o MedFound, treinado com grande corpus médico e registros clínicos e avaliado em doenças comuns, cenários externos e condições raras de oito especialidades. O trabalho incluiu comparação com médicos e avaliação de assistência diagnóstica, sustentando a viabilidade de uso como apoio. Ainda assim, a maior parte da evidência desse tipo continua concentrada em conjuntos de casos, bancos construídos para avaliação e estudos controlados. O salto para resultados clínicos depende de fatores que o benchmark não contém: prevalência, ruído, disponibilidade de exames, responsabilidade e consequências de seguir uma resposta errada.

Um modelo multimodal também pode errar de forma diferente de um sistema unimodal. Se texto e imagem apontam para conclusões incompatíveis, o modelo pode privilegiar a fonte errada sem deixar isso evidente. Pode produzir uma narrativa coerente que harmoniza artificialmente dados que deveriam permanecer em conflito até investigação adicional. Pode ainda receber uma imagem de baixa qualidade e compensar a incerteza visual com pistas presentes no texto, criando a impressão de que interpretou corretamente o exame. A integração de modalidades aumenta contexto, mas pode tornar a origem de uma conclusão mais difícil de reconstruir.

Por isso, rastreabilidade ganha valor. Uma resposta clínica útil deveria permitir reconhecer quais dados sustentaram determinada afirmação, quais estavam ausentes e onde existe incerteza. Em modelos capazes de trabalhar com prontuário, imagem e literatura, saber a proveniência do conteúdo é parte da possibilidade de revisão. Uma justificativa produzida em linguagem natural não é necessariamente uma reconstrução fiel do processo interno do modelo. O sistema pode explicar de maneira plausível depois de gerar uma resposta; essa explicação não deve ser tratada automaticamente como prova de causalidade nem como auditoria de seu raciocínio.

A orientação da Organização Mundial da Saúde para grandes modelos multimodais, publicada em 2024, responde precisamente a essa ampliação de poder.

O documento reconhece aplicações potenciais em cuidado, educação, pesquisa, desenvolvimento de medicamentos e tarefas administrativas, mas chama atenção para afirmações falsas ou incompletas, dados enviesados, automação indevida, riscos de segurança e desigualdade de acesso. O aspecto central é que modelos generalistas podem ser usados para finalidades que não foram previstas durante seu desenvolvimento. A flexibilidade que torna o sistema atraente também dificulta delimitar o perímetro de validação.

Essa dificuldade é ainda maior porque os modelos mudam. Versões são atualizadas, políticas de segurança são modificadas, componentes de recuperação podem ser trocados e fornecedores podem ajustar sistemas sem que o usuário perceba cada alteração. Um resultado obtido com determinada versão em um benchmark não deve ser tratado como propriedade eterna de um nome comercial. Para tecnologias generativas, registrar versão, configuração, contexto e data da avaliação deixa de ser detalhe técnico e passa a ser condição de interpretação da evidência.

Também é preciso preservar uma diferença conceitual: multimodal não significa completo. Um sistema pode receber texto e imagem e ainda desconhecer algo que o paciente não disse, que o prontuário não registrou ou que o exame não capturou. Pode integrar cinco fontes e continuar cego para condições sociais, preferências, barreiras de acesso e mudanças ocorridas fora do sistema de saúde. A sensação de visão abrangente é particularmente perigosa quando a interface reúne muitos dados em uma resposta elegante. Quantidade de modalidades não elimina ausência de contexto.

O ponto mais promissor talvez seja justamente usar esses modelos para reduzir fragmentação sem fingir que ela desapareceu. Uma boa ferramenta pode localizar informações dispersas, apontar contradições, sintetizar tendências e oferecer diferentes formatos de apresentação conforme a necessidade do usuário. Pode ajudar o radiologista a relacionar achados com exames anteriores, o clínico a navegar por anos de registros ou o paciente a compreender orientações já estabelecidas. O ganho está menos em produzir uma narrativa definitiva e mais em tornar a informação mais navegável sem ocultar sua origem.

Esse desenho exige que a saída generativa seja tratada como objeto clínico verificável. Afirmações factuais precisam poder ser conferidas; recomendações devem ser separadas de descrições; incerteza não pode desaparecer na edição; e a ausência de evidência deve ser representada como ausência, não preenchida com uma frase

plausível. Em outras palavras, o modelo precisa ser avaliado não apenas por aquilo que consegue dizer, mas por sua capacidade de não inventar quando não sabe.

É aqui que modelos generativos tornam visível uma questão que atravessou todos os capítulos anteriores. Na imagem, uma classificação errada pode ser contada como falso positivo ou falso negativo. Em um escore prognóstico, podemos medir calibração e discriminação. Em uma resposta aberta, o espaço de erro é muito maior: o sistema pode acertar o diagnóstico e errar a dose, omitir uma contraindicação, inventar uma fonte, atribuir um sintoma ao paciente errado ou oferecer uma justificativa convincente para uma recomendação inadequada. Avaliar apenas a resposta final empobrece a segurança.

A evolução recente da área mostra, portanto, uma tensão produtiva. Modelos de linguagem já conseguem executar tarefas clínicas complexas em estudos controlados; sistemas de sumarização e escribas ambientais já entraram em fluxos reais; modelos multimodais já combinam imagem e linguagem com flexibilidade impressionante. Ao mesmo tempo, os melhores estudos começam a mostrar que benefício não é propriedade abstrata de uma arquitetura. Ele depende da tarefa, do usuário, da integração, do comparador e do desfecho escolhido.

A pergunta que se abre, então, é mais exigente do que saber se a resposta parece boa. Como validar um sistema cuja saída é aberta, cuja versão pode mudar e cujo uso real talvez ultrapasse as tarefas previstas no estudo? Como medir erros raros, mas graves? Como comparar modelos que recebem contextos diferentes? Como testar segurança sem reduzir a clínica a um conjunto de perguntas artificiais? Quando a inteligência artificial passa a produzir linguagem, imagem interpretada, síntese e recomendação dentro do mesmo fluxo, a validação precisa acompanhar essa complexidade.

Essa pergunta desloca o foco da capacidade para a confiança. Depois de reconhecer o que sistemas conseguem classificar, prever, monitorar e gerar, desempenho técnico deixa de ser ponto de chegada e passa a ser apenas uma das evidências necessárias. Validação clínica, segurança do paciente, mudança de distribuição, monitoramento pós-implementação e detecção de falhas tornam-se partes do mesmo problema: demonstrar que uma tecnologia continua defensável quando sai do ambiente controlado e começa a produzir consequências reais.

CAPÍTULO 6
QUANDO DESEMPENHO NÃO BASTA: VALIDAÇÃO CLÍNICA, SEGURANÇA DO PACIENTE E MONITORAMENTO PÓS-IMPLEMENTAÇÃO

Validar é descobrir onde a promessa deixa de valer

Há um momento em que a pergunta sobre inteligência artificial em saúde muda de natureza. Até aqui, vimos sistemas capazes de reconhecer padrões, estimar riscos, acompanhar pacientes e produzir texto, sínteses e interpretações multimodais. Em cada uma dessas tarefas, o desempenho técnico importa. Mas, quando a tecnologia se aproxima do cuidado real, já não basta demonstrar que ela acerta em um conjunto de casos. É preciso descobrir em que condições continua acertando, como falha, quem percebe a falha e o que acontece com o paciente quando sua saída passa a participar de uma decisão.

Essa mudança parece semântica, mas define o sentido clínico da palavra validação. Um modelo não é validado em abstrato. Ele é avaliado para uma finalidade, uma população, um ambiente, uma versão e uma posição específica dentro do fluxo assistencial. Um algoritmo que funciona como segunda leitura de uma imagem enfrenta um problema de segurança diferente daquele que prioriza uma fila; um sistema que redige uma nota clínica não exige exatamente a mesma evidência de uma ferramenta que recomenda tratamento. Quanto maior a capacidade de uma saída alterar conduta e quanto mais difícil for detectar um erro antes que ele alcance o paciente, maior precisa ser a força da evidência que sustenta seu uso.

Por isso, a validação não deve ser imaginada como uma etapa única entre desenvolvimento e comercialização. Ela se parece mais com uma sequência de perguntas que se tornam progressivamente mais próximas da realidade. Primeiro, é necessário demonstrar que o modelo aprendeu algo que se mantém em dados que não participaram do treinamento. Depois, verificar se esse desempenho sobrevive a outras instituições, períodos, equipamentos ou populações. Em seguida, observar o sistema prospectivamente no ambiente em que será usado. Quando a ferramenta influencia decisões, torna-se necessário avaliar a interação com profissionais, o percurso assistencial e, conforme o risco, os efeitos clínicos produzidos por sua incorporação. Azad, Krumholz e Saria (2026) descrevem justamente a necessidade de sair de uma

cultura centrada em benchmarks para uma lógica de prontidão clínica construída por avaliação progressiva.

O primeiro risco aparece muito antes da implantação: confundir capacidade de ajuste com generalização. Modelos flexíveis conseguem aprender relações extremamente específicas de um conjunto de dados, inclusive características que não correspondem ao fenômeno clínico que se pretendia reconhecer. Se imagens de um hospital carregam marcas técnicas diferentes das de outro, se determinados exames são solicitados apenas para pacientes mais graves ou se o período de coleta coincide com uma prática institucional particular, o algoritmo pode utilizar essas pistas como atalhos. O resultado pode parecer excelente em um teste mal separado e deteriorar quando o atalho deixa de existir.

É por isso que independência entre desenvolvimento e avaliação é mais do que um detalhe metodológico. Dados do mesmo paciente, exames seriados ou padrões provenientes da mesma fonte podem criar uma proximidade artificial entre treinamento e teste. Da mesma forma, uma validação externa nominalmente realizada em outra base pode oferecer pouca informação se essa base reproduz quase exatamente o mesmo processo de aquisição. A pergunta relevante não é apenas se o arquivo veio de outro lugar, mas se a avaliação expôs o sistema a diferenças capazes de testar sua robustez.

Esse amadurecimento aparece nas próprias diretrizes metodológicas. TRIPOD+AI e PROBAST+AI, discutidos anteriormente, reforçaram transparência, risco de viés e aplicabilidade para modelos preditivos. Em 2025, o STARD-AI ampliou esse movimento para estudos de acurácia diagnóstica com inteligência artificial, incorporando itens relacionados a práticas de dados, descrição do teste baseado em IA, viés, equidade e generalização (SOUNDERAJAH et al., 2025). No mesmo ano, o TRIPOD-LLM propôs recomendações específicas para estudos com grandes modelos de linguagem, em resposta a uma dificuldade evidente: sistemas generativos podem variar conforme versão, instrução, contexto, estratégia de amostragem e forma de avaliação, tornando inadequado relatar apenas o nome comercial de um modelo como se ele identificasse uma intervenção estável (GALLIFANT et al., 2025).

Essas diretrizes não transformam um estudo fraco em evidência forte. Elas fazem algo anterior e indispensável: tornam visível o que foi feito, com quais dados, em qual versão, para qual finalidade e com quais limitações. Transparência é condição para

julgamento, não substituto de desenho adequado. Um artigo pode relatar perfeitamente um estudo retrospectivo e continuar incapaz de responder se a tecnologia melhora o cuidado. Da mesma forma, uma curva de desempenho pode ser metodologicamente correta e ainda medir um desfecho distante daquilo que interessa ao paciente.

A diferença fica clara quando passamos de modelos para intervenções. CONSORT-AI e SPIRIT-AI foram desenvolvidos para ensaios clínicos que incluem componentes de inteligência artificial e acrescentaram exigências próprias desse tipo de tecnologia: descrição do uso pretendido, requisitos de entrada, saída produzida, treinamento dos usuários, interação humano-IA e análise de erros, entre outros elementos (LIU et al., 2020; CRUZ RIVERA et al., 2020). O ponto central não é burocrático. Se o efeito depende de como um profissional interpreta um alerta ou incorpora uma recomendação, o algoritmo isolado não é a intervenção completa. O que precisa ser avaliado é o conjunto formado por modelo, interface, usuário e fluxo.

Isso também explica por que desfechos precisam acompanhar a posição da IA na cadeia do cuidado. Para um classificador de imagem, sensibilidade, especificidade, calibração e taxa de exames não interpretáveis podem ser importantes. Para um sistema de triagem, tempo até avaliação, sobrecarga, prioridade incorreta e casos perdidos passam a importar. Para um suporte terapêutico, entram adesão à recomendação, eventos adversos e desfechos clínicos. Para um escriba, a avaliação pode incluir tempo documental, qualidade da nota e erros clinicamente relevantes. A métrica deve seguir a consequência da ferramenta, não a conveniência do pesquisador.

A validação clínica, portanto, exige abandonar a ideia de que existe um número capaz de resumir segurança. Área sob a curva, sensibilidade, especificidade e acurácia continuam úteis, mas precisam ser interpretadas com calibração, prevalência, limiar de decisão, distribuição de erros, desempenho por subgrupos e efeito sobre o processo assistencial. Um sistema pode manter acurácia global e piorar exatamente na faixa de risco em que a decisão é tomada. Pode também melhorar um indicador intermediário e não alterar um desfecho relevante. A avaliação de uma tecnologia clínica começa quando deixamos de perguntar apenas quanto ela acerta e passamos a perguntar quais consequências produz quando acerta e quando erra.

Antes de interferir no cuidado, observar em silêncio

Existe uma etapa particularmente útil entre o estudo retrospectivo e a exposição direta de profissionais e pacientes à saída de um algoritmo: a avaliação prospectiva em

modo silencioso. Nessa configuração, o sistema é executado dentro do ambiente real, recebe dados como receberia após a implantação e produz resultados em tempo compatível com a rotina, mas essas saídas não são usadas para modificar o cuidado. A vantagem é testar o modelo contra a desordem da prática sem acrescentar, naquele momento, o risco de uma recomendação inadequada influenciar a conduta.

Essa fase consegue revelar problemas que bases retrospectivas frequentemente escondem. Dados podem chegar atrasados, campos podem estar vazios, unidades podem mudar, interfaces podem transformar valores, imagens podem apresentar qualidade pior do que a observada durante o desenvolvimento e integrações podem falhar. Também é possível descobrir que a frequência real de uso é diferente da imaginada ou que uma parcela relevante das entradas cai fora do domínio de treinamento. Um modelo pode ser estatisticamente competente e operacionalmente inviável porque depende de uma informação que só fica disponível depois do momento em que sua previsão teria valor.

A literatura recente mostra, entretanto, que ainda não existe uniformidade sobre o que exatamente constitui um bom estudo silencioso. Tikhomirov et al. (2026) revisaram 75 avaliações silenciosas publicadas entre 2015 e 2025 e encontraram grande heterogeneidade de terminologia, objetivos e práticas. Métricas como área sob a curva eram comuns, enquanto a verificação das saídas contra um padrão clínico em situação real e a análise de componentes sociotécnicos apareciam com muito menos consistência. A constatação é importante porque executar um algoritmo em segundo plano não garante, por si só, uma avaliação clínica abrangente. O modo silencioso é uma oportunidade de aproximar a validação do mundo real, mas precisa ser desenhado para responder perguntas que esse mundo realmente apresenta.

Um estudo publicado no JAMA em 2026 oferece um exemplo particularmente didático de avaliação em etapas. Zhang et al. desenvolveram um sistema baseado em tomografia de coerência óptica para apoiar a identificação de edema macular diabético. Antes de testar seu efeito sobre o encaminhamento, o sistema foi executado prospectivamente em modo silencioso em 603 pessoas com diabetes, totalizando 1.200 exames oculares. Nessa fase, 7,2% dos exames foram classificados como não graduáveis e, entre os graduáveis, 4,4% ficaram na faixa de incerteza. Para detecção de edema macular diabético, o sistema alcançou sensibilidade de 98,8% e especificidade de 90,7% (ZHANG et al., 2026).

O valor metodológico desse exemplo está justamente no fato de que a história não termina nesses números. Depois da fase silenciosa, os autores realizaram um ensaio randomizado multicêntrico com 276 participantes encaminhados por suspeita de edema macular diabético. No grupo de intervenção, oftalmologistas combinavam o relatório de rastreamento convencional com a saída do sistema de IA; no controle, o encaminhamento seguia o fluxo baseado no rastreamento convencional. A taxa de encaminhamentos falsamente positivos foi de 24,1% no grupo com IA e 69,1% no controle, cumprindo o critério de não inferioridade previamente estabelecido; a sensibilidade do encaminhamento foi de 100% nos dois grupos. Por razões éticas, todos os participantes realizaram avaliação especializada, de modo que as decisões de não encaminhar no grupo de intervenção foram efetivamente hipotéticas. Essa limitação impede transformar o ensaio em prova de desfechos assistenciais de longo prazo, mas não reduz sua contribuição para demonstrar uma estratégia de avaliação progressiva (ZHANG et al., 2026).

A sequência é mais informativa do que um estudo que pula diretamente do banco de dados para o uso clínico. A fase silenciosa permite testar entrada, qualidade, incerteza e desempenho no ambiente real; o ensaio seguinte pergunta o que muda quando o resultado entra no fluxo. Essa lógica é compatível com o DECIDE-AI, que propõe avaliação clínica inicial de sistemas de apoio à decisão em condições reais antes de estudos de maior escala, e com a noção de que a força da avaliação deve crescer junto com o risco e a influência da tecnologia sobre a conduta.

Ainda assim, estudos silenciosos possuem uma limitação estrutural: eles não conseguem observar completamente o que acontece quando seres humanos veem a saída. Um alerta oculto não causa fadiga de alerta. Uma recomendação que ninguém recebe não gera ancoragem, contestação ou dependência. Uma interface não utilizada não revela se o profissional compreende a incerteza. Por isso, o sucesso em modo silencioso não demonstra automaticamente sucesso em uso aberto. Ele reduz parte da incerteza técnica e operacional, mas não substitui a avaliação da interação humano-máquina.

O mesmo vale para ferramentas generativas. É possível testar silenciosamente um sistema que produz resumos ou rascunhos e comparar suas saídas com documentos humanos, mas a implantação altera comportamento. Profissionais podem revisar de maneira superficial um texto fluente, gastar mais tempo corrigindo certas classes de erro ou mudar a forma de registrar informações porque sabem que um modelo fará a síntese. A intervenção modifica o ambiente que deveria apenas servir de cenário para

sua avaliação. Esse efeito de retroalimentação é uma razão adicional para combinar métricas técnicas com observação de uso.

Ensaios randomizados continuam importantes porque permitem comparar estratégias de cuidado com maior proteção contra confundimento. Mas nem toda pergunta exige mortalidade como desfecho e nem toda ferramenta precisa do mesmo desenho. O estudo deve ser proporcional ao risco e à função. Uma tecnologia administrativa de baixo impacto pode ser avaliada por qualidade, eficiência e carga de trabalho; um sistema que autonomamente retém um encaminhamento, modifica dose ou prioriza pacientes instáveis exige uma justificativa muito mais forte de segurança e benefício. O rigor não significa usar sempre o desenho mais caro. Significa não pedir a uma evidência fraca que sustente uma decisão maior do que ela consegue suportar.

Essa proporcionalidade também ajuda a evitar o problema oposto: exigir de toda inovação uma prova impossível antes de qualquer aprendizagem no mundo real. O caminho mais seguro não é escolher entre "liberar" ou "proibir" com base em um único estudo, mas construir etapas em que o risco seja controlado enquanto a evidência se aproxima progressivamente da prática. O consenso FUTURE-AI, publicado em 2025 a partir de um grupo internacional com 117 especialistas de 50 países, organiza essa visão em princípios de equidade, universalidade, rastreabilidade, usabilidade, robustez e explicabilidade e enfatiza avaliação e mitigação de risco ao longo de todo o ciclo de vida (LEKADIR et al., 2025).

Depois da implantação, segurança vira processo

A aprovação ou a decisão institucional de implantar um sistema não congela o mundo ao redor dele. Populações mudam, prevalências mudam, equipamentos são substituídos, protocolos são atualizados, profissionais alteram comportamento e o próprio software pode receber novas versões. Em modelos de aprendizado, essa instabilidade não é uma exceção periférica; é parte do ambiente em que a tecnologia opera. Por isso, a validação realizada antes da implantação informa o início do uso, mas não garante que o desempenho permanecerá igual meses ou anos depois.

O termo drift costuma ser usado para descrever mudanças que alteram a relação entre o sistema e os dados que recebe. Pode haver mudança na distribuição das entradas, na frequência dos desfechos ou na relação entre preditores e aquilo que se deseja antecipar. Na prática, isso pode acontecer de maneira banal: um novo equipamento produz imagens com características diferentes; uma política de internação muda o perfil dos pacientes; um teste laboratorial troca de método; uma

pandemia altera prevalências, fluxos e decisões. O algoritmo continua executando o mesmo código, mas o mundo que alimenta esse código já não é o mesmo.

Um estudo de 2025 na Veterans Health Administration mostra como essa deterioração pode ser silenciosa mesmo em grande escala. Kolla et al. avaliaram o algoritmo CAN, utilizado nacionalmente para estratificação de risco, em mais de 27 milhões de observações correspondentes a aproximadamente 7,2 milhões de veteranos entre 2016 e 2021. Nesse período, o valor preditivo positivo caiu cerca de quatro pontos percentuais e a taxa de falsos positivos aumentou, correspondendo a mais de 18 mil resultados falsamente positivos adicionais, enquanto algumas métricas agregadas, como acurácia, permaneceram estáveis. Os autores também observaram mudanças em prevalência de desfechos e em variáveis demográficas, laboratoriais e de utilização dos serviços (KOLLA et al., 2025).

Esse exemplo desmonta a ideia de que monitorar uma única métrica seja suficiente. Se a acurácia global continua parecida, uma instituição pode concluir que nada mudou, enquanto o valor operacional do escore se deteriora exatamente no grupo para o qual recursos são direcionados. Em sistemas de saúde, a consequência do drift depende do uso. Uma queda de valor preditivo positivo pode significar mais avaliações desnecessárias; uma mudança de calibração pode fazer probabilidades parecerem mais altas ou mais baixas do que o risco real; uma piora concentrada em um subgrupo pode permanecer invisível na média.

O monitoramento também não pode depender apenas de esperar que desfechos ocorram para recalcular desempenho. Em algumas aplicações, o padrão de referência aparece dias ou meses depois; em outras, obter rótulos exige revisão humana. Kore et al. (2024) estudaram 239.235 radiografias de tórax obtidas antes e depois da emergência da COVID-19 e mostraram que a mudança na distribuição dos dados podia ser detectada por métodos voltados às entradas mesmo quando a área sob a curva permanecia relativamente estável. A lição é sutil: ausência de queda imediata em uma métrica agregada não significa ausência de mudança relevante no ambiente.

Por isso, um programa de monitoramento maduro precisa observar várias camadas ao mesmo tempo. A primeira é técnica: falhas de integração, dados ausentes, latência, mudanças de equipamento, proporção de entradas rejeitadas e distribuição das variáveis. A segunda é estatística e clínica: calibração, sensibilidade, especificidade, valores preditivos, erros por classe e desempenho nos grupos em que o sistema é efetivamente utilizado. A terceira é operacional: quantos alertas chegam, quantos são

ignorados, que condutas são tomadas, se houve mudança na carga de trabalho e se o fluxo criado pelo algoritmo produz gargalos em outro ponto do serviço. A quarta é de segurança: incidentes, quase incidentes, reclamações e padrões inesperados que uma métrica de desempenho não consegue antecipar sozinha.

Monitorar, porém, só é útil quando existe uma regra sobre o que fazer diante do sinal. Uma instituição precisa definir quais mudanças exigem investigação, quando o sistema deve ser recalibrado, quando uma atualização precisa de nova validação, quando usuários devem ser avisados e em que circunstâncias o modelo deve ser temporariamente suspenso. Caso contrário, o painel de monitoramento se transforma em um arquivo de anomalias sem consequência. Segurança não é detectar que algo mudou; é possuir capacidade institucional para responder de maneira proporcional à mudança.

Essa lógica já aparece de forma explícita em orientações regulatórias recentes. O documento final de Good Machine Learning Practice do International Medical Device Regulators Forum, publicado em 2025, organiza o desenvolvimento de dispositivos baseados em aprendizado de máquina a partir de uma perspectiva de ciclo de vida, incluindo uso pretendido, qualidade de software e dados, avaliação do desempenho da equipe humano-IA, testes apropriados e monitoramento de modelos implantados (IMDRF, 2025). Nos Estados Unidos, a orientação final da FDA sobre Predetermined Change Control Plans, publicada em agosto de 2025, estabelece uma estrutura para que determinadas modificações futuras de funções de software habilitadas por IA sejam previamente descritas, acompanhadas de metodologia de desenvolvimento, validação e avaliação de impacto (FDA, 2025).

A existência de um plano de mudança não significa autorização para modificar indefinidamente um sistema sem evidência. Pelo contrário, reconhece que atualização também é uma intervenção sobre a tecnologia validada. Uma nova versão pode corrigir um problema e introduzir outro, melhorar desempenho global e piorar um subgrupo, ou alterar a forma como usuários interpretam a saída. O objeto avaliado deixa de ser apenas "o algoritmo" e passa a incluir sua história de versões. Em sistemas generativos, esse cuidado é ainda mais importante porque fornecedores podem alterar modelos de base, filtros, recuperação de informação e comportamento de resposta em ciclos muito mais rápidos do que os tradicionais.

Van Amsterdam et al. (2026) chamam atenção para um desafio que tende a crescer: ensaios de sistemas continuamente monitorados e atualizados precisam

distinguir mudanças que fazem parte da própria intervenção de mudanças relacionadas à supervisão do estudo. Se um modelo é atualizado durante seu uso, a pergunta "qual versão produziu este resultado?" deixa de ser administrativa. Ela se torna necessária para atribuir efeito, reproduzir evidência e compreender eventos adversos. A rastreabilidade de versão, dados, limiar e configuração passa a ser parte da segurança clínica.

Isso conduz a uma conclusão importante: segurança de IA não pode ser terceirizada para o profissional individual que recebe a saída. O médico ou outro profissional precisa manter julgamento crítico, mas não controla a procedência dos dados, a arquitetura do modelo, a política de atualização, a integração do sistema ou o contrato com o fornecedor. Instituições precisam saber quais ferramentas estão ativas, para quais finalidades, em quais setores, com que versão e sob quais indicadores de monitoramento. Desenvolvedores precisam fornecer informação suficiente para que falhas sejam investigadas. Reguladores precisam definir expectativas proporcionais ao risco. Responsabilidade humana continua existindo, mas precisa ser sustentada por governança capaz de tornar o sistema auditável.

A melhor analogia talvez não seja pensar em uma inteligência artificial como um exame que "passou" ou "não passou" em validação. É mais útil tratá-la como um componente de um sistema clínico vivo, cuja segurança depende de manter coerência entre finalidade, população, desempenho e resposta institucional. Um modelo confiável hoje pode deixar de ser adequado amanhã sem que seu código esteja corrompido. O hospital também muda, a epidemiologia muda e a própria decisão que o algoritmo influencia pode mudar. A confiança, portanto, não é uma propriedade permanente adquirida no lançamento; é uma relação que precisa ser continuamente sustentada por evidência.

Esse entendimento também muda o significado de erro. Antes da implantação, o erro costuma aparecer como uma célula em uma matriz de confusão. Depois da implantação, ele ganha trajetória: alguém recebeu uma saída, interpretou-a, tomou ou deixou de tomar uma decisão e essa decisão encontrou um paciente com características específicas. Investigar segurança exige reconstruir essa trajetória. Às vezes o problema estará no modelo; em outras, na entrada, na interface, no limiar, no treinamento, na sobrecarga ou na forma como o serviço incorporou a ferramenta. Chamar tudo de "erro da IA" pode ser tão pouco informativo quanto chamar tudo de "erro humano".

FUTURE-AI e as orientações de ciclo de vida convergem justamente nesse ponto: robustez e confiabilidade não são atributos demonstrados uma vez, mas práticas de avaliação, rastreabilidade e monitoramento que acompanham a tecnologia enquanto ela permanece em uso (LEKADIR et al., 2025; IMDRF, 2025). A inteligência artificial pode ampliar a capacidade de detectar sinais precoces, organizar informação e reduzir determinadas formas de variação. Mas o benefício só permanece defensável se a organização também estiver preparada para detectar quando a ferramenta deixa de cumprir a função para a qual foi aceita.

Há, contudo, uma pergunta que o desempenho global ainda não responde. Mesmo quando uma tecnologia mantém suas métricas ao longo do tempo, esse resultado pode esconder diferenças relevantes entre pessoas. Um sistema pode ser estável na média e sistematicamente menos sensível em determinada população; pode produzir mais falsos positivos em um grupo; pode usar como preditores variáveis que carregam desigualdades de acesso; pode ser atualizado com dados que reforçam uma distribuição já excludente. Monitorar segurança sem olhar para quem recebe benefício e quem concentra erro deixa uma parte essencial do risco invisível.

A confiabilidade de um sistema, portanto, não pode ser declarada apenas em termos globais. Mesmo quando o desempenho se mantém ao longo do tempo, permanece uma pergunta que a média não responde: para quem ele funciona, quem fica de fora e quem concentra os erros? Sem essa leitura, a automação pode preservar desigualdades históricas justamente porque as apresenta sob a aparência neutra de uma métrica agregada.

CAPÍTULO 7
QUEM FICA DE FORA DOS DADOS? VIÉS ALGORÍTMICO, EQUIDADE E DESIGUALDADES EM SAÚDE

O viés começa antes do modelo

No capítulo anterior, a confiança em inteligência artificial foi tratada como uma relação que precisa sobreviver à mudança: novos pacientes, novas rotinas, novas versões e novas condições de uso. Há, porém, um problema que pode permanecer invisível mesmo quando as métricas globais parecem estáveis. Um modelo pode continuar funcionando bem na média e, ainda assim, distribuir seus erros de forma desigual. Pode reconhecer uma doença com maior facilidade em determinados grupos, priorizar algumas pessoas antes de outras ou transformar uma diferença histórica de acesso em uma regra matemática aparentemente neutra. Nesse ponto, avaliar segurança exige perguntar não apenas se a tecnologia funciona, mas para quem funciona e quem suporta o custo quando ela falha.

A palavra viés costuma sugerir uma distorção introduzida pelo algoritmo. Essa imagem é insuficiente. Em saúde, desigualdades podem entrar muito antes do treinamento. Elas aparecem em quem consegue chegar ao serviço, em quais exames são solicitados, em quem recebe determinado diagnóstico, na qualidade do registro, no equipamento disponível, no tempo dedicado à consulta e nas decisões que produziram os desfechos usados como rótulo. O conjunto de dados não é uma amostra neutra do mundo; ele é uma sedimentação de práticas clínicas, barreiras de acesso e escolhas institucionais. Quando um modelo aprende com esse material, pode aprender também a organização desigual do cuidado.

Isso torna especialmente importante distinguir ausência de representação de representação inadequada. Um grupo pode estar pouco presente no banco e, por isso, oferecer poucos exemplos para que o modelo aprenda. Mas também pode estar numericamente presente e ainda ser mal representado se os dados disponíveis forem sistematicamente menos completos, se os rótulos forem menos confiáveis ou se a relação entre sinais e desfechos for diferente naquele contexto. Contar pessoas é necessário, mas não suficiente. Representatividade clínica envolve qualidade, diversidade de situações, prevalência, trajetória assistencial e adequação daquilo que foi medido.

O problema do alvo é ainda mais profundo. No Capítulo 4, vimos o caso descrito por Obermeyer et al. (2019), em que um algoritmo utilizado para selecionar pacientes para programas adicionais de cuidado previa gastos em saúde como aproximação de necessidade clínica. A escolha parecia objetiva e produzia boa capacidade preditiva para aquilo que havia sido definido como alvo. O erro estava justamente na definição do que deveria ser previsto. Como pacientes negros historicamente recebiam menos recursos para níveis semelhantes de doença, gasto menor não significava menor necessidade. O algoritmo aprendeu com fidelidade uma variável que já carregava desigualdade e, ao fazê-lo, reduziu a probabilidade de encaminhar pacientes negros para ajuda adicional.

Esse episódio é importante porque desloca a discussão. Um modelo pode não utilizar raça como entrada e ainda produzir disparidade racial. Pode não conter uma regra discriminatória explícita e ainda assim reproduzir uma estrutura desigual porque utiliza um substituto que se comporta de modo diferente entre grupos. Custo, frequência de consultas, número de exames, adesão registrada ou presença de determinado código diagnóstico podem parecer medidas objetivas, mas todas dependem de oportunidades anteriores de acesso e cuidado. Em aprendizado de máquina, neutralidade da variável não garante neutralidade do significado.

Tampouco basta retirar atributos sensíveis da base. Em 2022, Gichoya et al. demonstraram que modelos de deep learning conseguiam reconhecer a identidade racial autorreferida de pacientes a partir de diferentes modalidades de imagem médica, incluindo radiografias de tórax, mamografias e tomografias, mesmo quando especialistas humanos não dispunham de um marcador visual equivalente para realizar essa identificação. O estudo não demonstrou que raça seja uma propriedade biológica contida na imagem; demonstrou que os modelos capturam sinais correlacionados com categorias demográficas de maneiras que ainda não compreendemos completamente. Eliminar uma coluna chamada “raça” do banco, portanto, não impede que outras características codifiquem informações relacionadas a ela (GICHOYA et al., 2022).

Essa distinção precisa ser preservada com cuidado. Categorias raciais e étnicas são construídas histórica e socialmente e variam entre contextos. Utilizá-las para auditar disparidades não significa tratá-las como causas biológicas simples. Em muitos casos, elas funcionam como marcadores de exposições, desigualdades, experiências sociais e padrões de acesso que a medicina precisa compreender sem essencializar. O desafio é duplo: ignorar essas categorias pode esconder dano; usá-las de maneira biologizante pode criar outra forma de erro.

O mesmo vale para sexo, idade, deficiência, condição socioeconômica, idioma e território. Cada variável pode ter significado diferente conforme a tarefa. Em determinada aplicação, diferenças por idade podem refletir manifestação clínica legítima; em outra, podem indicar que o modelo encontrou um atalho relacionado ao equipamento ou ao padrão de atendimento. Equidade não nasce de proibir qualquer diferença entre grupos. Nasce de investigar se a diferença observada possui justificativa clínica, se produz benefício ou dano desigual e se permanece aceitável quando o sistema é usado fora do ambiente em que foi desenvolvido.

Esse raciocínio também impede uma conclusão comum e equivocada: a de que um algoritmo seria automaticamente mais imparcial do que uma decisão humana porque aplica a mesma função matemática a todos. Aplicar a mesma regra a entradas produzidas por condições diferentes pode preservar desigualdades. Ao mesmo tempo, a inteligência artificial não está condenada a ampliá-las. Como veremos, alguns modelos conseguem reduzir diferenças de desempenho quando dados, objetivo, treinamento e avaliação são desenhados explicitamente para isso. A tecnologia não possui uma direção moral própria; o que importa é como o problema foi formulado e quais consequências são aceitas como parte de sua solução.

A média pode esconder quem concentra o erro

A forma mais simples de perder uma desigualdade é olhar apenas para a média. Imagine um sistema com excelente sensibilidade global porque funciona muito bem no grupo que representa a maior parte dos exames, mas perde uma proporção significativamente maior de casos em uma população menor. A métrica agregada continuará atraente. Para o paciente pertencente ao grupo em que os falsos negativos se concentram, porém, a média não oferece proteção. Auditoria de equidade começa justamente por decompor desempenho e perguntar onde os erros se acumulam.

Seyyed-Kalantari et al. (2021) mostraram esse problema em classificadores de radiografias de tórax treinados com grandes bases públicas. Os autores analisaram subgrupos definidos por sexo, raça/etnia, idade e tipo de seguro, utilizado como indicador imperfeito de condição socioeconômica. Os modelos apresentaram padrões de subdiagnóstico mais frequentes em populações historicamente subatendidas, incluindo mulheres, pacientes negros, pacientes hispânicos, pessoas mais jovens e beneficiários do Medicaid em determinados conjuntos. Quando atributos eram combinados, algumas populações interseccionais apresentavam risco ainda maior de serem classificadas incorretamente como sem achados.

A interseccionalidade é decisiva porque categorias analisadas isoladamente podem produzir falsa tranquilidade. Um modelo pode ter desempenho aparentemente semelhante entre homens e mulheres e ainda falhar de modo desproporcional entre mulheres de determinado grupo racial ou faixa etária. Do mesmo modo, uma análise por raça pode esconder diferenças importantes associadas a sexo, idade, condição socioeconômica ou combinação desses fatores. Quanto mais estratificada a avaliação, entretanto, menor tende a ser o número de casos em cada célula e maior a incerteza estatística. A resposta não é abandonar a análise, mas reconhecer seus limites e planejar conjuntos de validação capazes de sustentar perguntas de equidade relevantes.

Dermatologia oferece um exemplo particularmente intuitivo porque diferenças de representação podem ser percebidas visualmente. Em um experimento internacional publicado em 2024, Groh et al. avaliaram dermatologistas, médicos da atenção primária e outros profissionais diante de imagens de doenças de pele em diferentes tonalidades. Especialistas e generalistas apresentaram menor acurácia em imagens de pele mais escura. O suporte de um sistema de deep learning desenhado com preocupação de equidade melhorou a acurácia de ambos os grupos em mais de 33%, mas não eliminou automaticamente as disparidades: entre médicos generalistas, a diferença de desempenho entre tonalidades chegou a aumentar (GROH et al., 2024).

Esse resultado é valioso porque desmonta duas expectativas ao mesmo tempo. A primeira é que a IA necessariamente amplifica qualquer viés existente. Não necessariamente: o suporte algorítmico elevou o desempenho global de forma importante. A segunda é que melhorar a acurácia global corrige automaticamente desigualdades. Também não. Uma intervenção pode beneficiar todos e ainda beneficiar alguns mais do que outros. Em termos clínicos, “melhor” e “mais equitativo” são propriedades relacionadas, mas não idênticas.

O desafio aumenta quando o sistema muda de hospital ou população. Yang et al. (2024) examinaram milhares de modelos de imagem em radiologia, dermatologia e oftalmologia e mostraram que métodos capazes de reduzir diferenças de desempenho no conjunto em que foram desenvolvidos nem sempre preservavam a mesma equidade em bases externas. Em outras palavras, justiça local não garantia justiça depois de uma mudança de distribuição. Os autores também observaram que modelos que codificavam menos informação demográfica em suas representações podiam apresentar melhor comportamento de equidade em ambientes externos do que aqueles escolhidos apenas por apresentar a menor diferença entre grupos no conjunto original.

Esse achado acrescenta uma camada importante ao capítulo anterior. O drift não ameaça apenas a acurácia global; pode alterar grupos de maneira assimétrica. Um modelo pode piorar para todos quando muda de instituição, mas piorar muito mais para um grupo específico. Se o monitoramento acompanhar apenas desempenho agregado, a deterioração distributiva permanecerá invisível. Segurança e equidade, portanto, não podem ser governadas em trilhas separadas. O mesmo sistema de vigilância pós-implementação que acompanha calibração, sensibilidade e taxa de alertas precisa também observar como essas métricas se comportam entre populações clinicamente relevantes.

A escolha da métrica de equidade também precisa responder ao dano que se pretende evitar. Em um rastreamento em que perder doença é especialmente grave, diferenças de falsos negativos podem ser mais relevantes do que diferenças de acurácia total. Em uma ferramenta que prioriza procedimentos invasivos, desigualdade de falsos positivos pode expor um grupo a mais intervenções desnecessárias. Em modelos probabilísticos, calibração por grupo pode ser essencial se o mesmo valor de risco estiver associado a frequências diferentes do desfecho. Não existe uma métrica universal capaz de representar todas as concepções de justiça ao mesmo tempo.

Esse último ponto impede que equidade seja reduzida a uma operação matemática de "igualar números". Algumas métricas de justiça podem entrar em tensão entre si, especialmente quando prevalências e riscos basais diferem entre grupos. Forçar igualdade em um indicador pode piorar outro clinicamente relevante. A decisão, portanto, precisa começar pelo percurso do paciente: qual erro cria atraso, tratamento excessivo, exclusão de uma fila ou redução de acesso? Só depois dessa pergunta faz sentido escolher a medida que permitirá detectar a distribuição desse dano.

Até mesmo explicações podem alterar essa distribuição. Em estudo publicado em agosto de 2026, Xu et al. avaliaram pessoas leigas e médicos da atenção primária usando um sistema de apoio dermatológico treinado com restrições de equidade. O modelo reduziu disparidades entre tonalidades de pele e melhorou a acurácia. Porém, a maneira de apresentar explicações e a ordem da interação modificaram o comportamento humano: usuários leigos demonstraram maior viés de automação diante de explicações geradas por modelos de linguagem quando a IA errava, enquanto médicos experientes mostraram maior resistência a esse efeito. Apresentar a recomendação algorítmica antes da decisão humana também aumentou preocupação com ancoragem (XU et al., 2026).

Isso mostra que uma ferramenta pode ser "justa" em seu desempenho isolado e produzir efeitos diferentes depois que encontra usuários. Se determinados profissionais possuem menos experiência com uma apresentação clínica, se pacientes interpretam a linguagem automatizada de maneiras distintas ou se uma interface conduz grupos diferentes a decisões diferentes, a auditoria não pode terminar na saída do modelo. Equidade precisa ser observada no sistema humano-tecnológico, inclusive na forma como confiança, explicação e autoridade são distribuídas.

Equidade não é um ajuste final

Uma estratégia séria de equidade começa antes do treinamento e continua depois da implantação. No início, isso significa perguntar quem participou da definição do problema. Uma ferramenta para reduzir faltas a consultas pode parecer bem sucedida se prevê com alta precisão quem não comparecerá, mas sua utilidade muda completamente se a intervenção resultante for cancelar vagas antecipadamente em vez de remover barreiras de transporte, comunicação ou horário. O mesmo modelo pode apoiar cuidado ou exclusão dependendo da decisão que será acoplada à previsão.

Participação importa justamente porque pesquisadores, fornecedores e gestores não enxergam todas as formas de dano a partir de seus próprios lugares. A orientação da Organização Mundial da Saúde para inteligência artificial em saúde coloca inclusão e equidade entre seus princípios centrais e recomenda que pessoas afetadas pelas tecnologias participem de seu desenho, desenvolvimento e avaliação (OMS, 2021). Em sistemas generativos, essa necessidade se amplia porque falhas podem aparecer em linguagem, estereótipos, omissões, pressupostos culturais e recomendações cuja inadequação não é capturada por um teste de acurácia convencional.

Pfohl et al. (2024) partiram exatamente dessa dificuldade ao propor uma estrutura e conjuntos de avaliação voltados a danos de equidade em respostas longas produzidas por grandes modelos de linguagem. Em vez de procurar apenas respostas factualmente falsas, os autores examinaram dimensões como estereótipos, exclusão, qualidade desigual de informação, possibilidade de dano e diferenças de utilidade. A lógica é importante: um modelo pode fornecer uma resposta tecnicamente correta e ainda ser menos acionável, menos respeitosa ou mais propensa a pressupostos inadequados quando a pergunta envolve determinadas identidades ou contextos.

A evidência experimental reforça essa preocupação. Omar et al. (2025) avaliaram nove grandes modelos de linguagem em mais de 1,7 milhão de saídas derivadas de mil casos de emergência apresentados em múltiplas versões sociodemográficas, mantendo

constantes os detalhes clínicos. A simples alteração de atributos sociais modificou recomendações em parte dos modelos. Casos identificados como pacientes negros, pessoas sem moradia ou pessoas LGBTQIA+ foram, em determinadas comparações, direcionados com maior frequência a cuidados urgentes, intervenções invasivas ou avaliações de saúde mental sem que a informação clínica justificasse necessariamente essa mudança. O estudo não significa que qualquer menção a um atributo social deva ser ignorada; mostra que sua influência precisa ser clinicamente justificável e auditável.

Essa distinção é particularmente importante porque contexto social pode ser clinicamente relevante. Moradia instável pode afetar segurança de alta, capacidade de armazenar medicamentos ou possibilidade de acompanhamento. Idioma pode modificar a forma de comunicação e a necessidade de intérprete. Sexo pode influenciar epidemiologia de determinadas condições. A solução não é produzir um modelo cego a qualquer diferença humana. É impedir que categorias sociais funcionem como atalhos para recomendações que não derivam de necessidades clínicas ou objetivos explicitamente defensáveis.

Mitigação, por isso, não pode ser reduzida a uma única técnica. Em alguns casos, melhorar representação e qualidade dos dados será essencial. Em outros, será necessário revisar o alvo, corrigir rótulos, recalibrar limiares, escolher outro modelo, redesenhar a intervenção ou admitir que a aplicação não deve ser utilizada em determinada população. Métodos de reponderação, balanceamento, treinamento adversarial e restrições de equidade podem reduzir diferenças em conjuntos específicos, mas os resultados de Yang et al. (2024) lembram que uma solução otimizada localmente pode falhar quando o contexto muda. Nenhum procedimento de debiasing elimina a necessidade de validação externa.

Também existe risco em “corrigir” o modelo sem corrigir o sistema. Se uma tecnologia passa a identificar com maior precisão pessoas que precisam de cuidado, mas a rede continua sem capacidade de oferecer esse cuidado, a desigualdade pode apenas se deslocar para a etapa seguinte. Se um algoritmo reduz subdiagnóstico em um grupo, mas o encaminhamento subsequente enfrenta fila maior, barreiras financeiras ou distância geográfica, a melhoria computacional não garante equidade assistencial. O desfecho de interesse precisa continuar sendo o percurso real, não apenas a distribuição das probabilidades produzidas na tela.

Por isso, avaliações de impacto deveriam observar não somente quem recebe uma classificação, mas quem efetivamente recebe benefício. Em uma ferramenta de

priorização, interessa saber quem foi chamado, quem conseguiu comparecer, quem completou a investigação e quem recebeu tratamento oportuno. Em um sistema de comunicação, importa verificar compreensão e capacidade de agir sobre a informação. Em um modelo generativo, é necessário estudar se diferentes populações recebem respostas equivalentes em qualidade, segurança e adequação. Equidade é uma propriedade do caminho completo entre dado e consequência.

Essa abordagem exige indicadores definidos antes da implantação. A instituição precisa saber quais grupos são clinicamente relevantes para auditoria, quais métricas serão acompanhadas, qual diferença acionará investigação e quem terá autoridade para modificar limiares, suspender uso ou exigir nova validação. Diretrizes contemporâneas de boa prática e confiança em inteligência artificial médica aproximam esse tema da gestão de ciclo de vida, com atenção à representatividade da população, ao desempenho em condições de uso e ao monitoramento de riscos depois da implantação (IMDRF, 2025; LEKADIR et al., 2025). O ponto não é transformar cada hospital em laboratório de ciência de dados, mas impedir que disparidades previsíveis permaneçam sem dono.

Há uma tensão delicada nesse processo. Para descobrir se um sistema prejudica mais determinado grupo, é preciso muitas vezes saber a qual grupo cada pessoa pertence. Auditar desigualdade pode exigir dados sobre raça/etnia autorreferida, sexo, idade, deficiência, idioma, território ou condição socioeconômica. Mas coletar e manter essas informações aumenta obrigações de confidencialidade, segurança e governança. Se os atributos são eliminados em nome da privacidade, certas formas de discriminação tornam-se impossíveis de medir; se são coletados sem proteção adequada, o esforço de equidade pode criar novos riscos.

Esse paradoxo revela por que os capítulos seguintes não tratam de temas independentes. Equidade depende de dados suficientemente detalhados para tornar diferenças visíveis; privacidade exige limitar exposição, acesso e uso indevido desses mesmos dados. Explicabilidade pode ajudar a investigar por que um sistema produz resultados distintos, mas também pode criar uma falsa sensação de compreensão. Transparência é necessária para responsabilização, porém não autoriza divulgar informações que identifiquem pacientes ou revelem infraestrutura sensível. A governança da IA em saúde é justamente o trabalho de equilibrar valores legítimos que nem sempre apontam para a mesma direção.

A conclusão mais importante, portanto, não é que algoritmos sejam inevitavelmente enviesados nem que exista uma técnica capaz de torná-los perfeitamente justos. É que equidade precisa ser tratada como uma pergunta clínica e institucional contínua: quem foi representado, o que foi usado como verdade, qual erro importa, em quem ele ocorre, como muda quando o contexto muda e que consequência produz depois da saída do modelo. Quando essas perguntas são incorporadas ao desenvolvimento e ao monitoramento, a inteligência artificial pode inclusive ajudar a revelar desigualdades que antes estavam diluídas na rotina. Quando são ignoradas, a automação pode conferir velocidade e aparência de objetividade a padrões históricos que nunca foram neutros.

Esse conflito torna equidade e proteção de dados inseparáveis. Auditar diferenças exige, muitas vezes, olhar para atributos pessoais e trajetórias de cuidado; proteger pacientes exige limitar quem pode acessar essas informações, como elas circulam, para que são reutilizadas e por quanto tempo permanecem disponíveis. A utilidade clínica dos dados, portanto, precisa conviver com limites de acesso, transparência apropriada, explicabilidade e segurança sem fazer desaparecer justamente as diferenças que precisam ser medidas.

CAPÍTULO 8
DADOS QUE CUIDAM E DADOS QUE EXPÕEM: PRIVACIDADE, TRANSPARÊNCIA, EXPLICABILIDADE E CIBERSEGURANÇA

Quando o dado necessário também cria exposição

O capítulo anterior terminou diante de um paradoxo. Para descobrir se um sistema produz resultados desiguais, muitas vezes é necessário conhecer atributos que a própria proteção de dados exige tratar com cuidado: raça ou etnia, idade, sexo, território, deficiência, condições clínicas, trajetória assistencial. Quanto mais detalhada é a informação, maior pode ser a capacidade de reconhecer uma disparidade; ao mesmo tempo, maior pode ser o dano se essa informação for acessada, combinada ou utilizada fora de sua finalidade. Privacidade e equidade, portanto, não ocupam lados opostos. Elas se encontram no mesmo problema: decidir quais dados são necessários para cuidar, quem pode utilizá-los, por quanto tempo e sob quais garantias.

Na saúde, esse problema é especialmente sensível porque os dados não descrevem apenas uma consulta. Um prontuário longitudinal pode revelar doenças atuais e passadas, uso de medicamentos, hábitos, histórico familiar, saúde mental, sexualidade, exames genéticos, imagens e padrões de utilização do sistema. Informações aparentemente banais ganham outro significado quando combinadas. Um horário recorrente de atendimento, uma sequência de exames ou a frequência com que alguém procura determinado serviço podem funcionar como pistas sobre uma condição que nunca foi declarada diretamente. É por isso que a Lei Geral de Proteção de Dados Pessoais brasileira classifica dados referentes à saúde, além de dados genéticos e biométricos vinculados a uma pessoa, entre os dados pessoais sensíveis (BRASIL, 2018).

Privacidade, nesse contexto, não significa manter todo dado em segredo. A medicina depende de circulação legítima de informação: equipes precisam compartilhar resultados, serviços precisam coordenar cuidado, pesquisadores precisam produzir conhecimento e sistemas públicos precisam organizar políticas. O problema é perder a relação entre uso e finalidade. Price e Cohen (2019) chamam atenção para esse ponto ao discutir como a expansão de grandes bases médicas

modifica o alcance da privacidade. Uma informação coletada para tratar uma pessoa pode adquirir valor para treinamento de modelos, pesquisa, gestão, previsão de risco ou desenvolvimento comercial. Cada reutilização pode ser defensável, mas nenhuma deveria ser considerada automaticamente equivalente à finalidade original.

A inteligência artificial amplia essa tensão porque transforma o dado em matéria-prima de várias etapas. Antes que um modelo seja usado, informações podem ser extraídas, limpas, rotuladas, copiadas para ambientes de desenvolvimento, combinadas com outras bases e processadas por diferentes equipes ou fornecedores. Depois da implantação, surgem novos registros: entradas enviadas ao sistema, escores, saídas, correções humanas, logs de uso e dados de monitoramento. Em aplicações generativas, uma única interação pode conter trechos extensos do prontuário ou detalhes suficientes para identificar um paciente. A proteção precisa acompanhar todo esse percurso. Concentrar atenção apenas no banco de dados original deixa invisível uma parte importante da superfície de exposição (OMS, 2024).

Anonimizar dados é uma das respostas mais intuitivas, mas não encerra o problema. Remover nome, CPF ou endereço reduz risco, porém conjuntos ricos podem conter combinações que tornam uma pessoa novamente distinguível quando cruzadas com outras fontes. O desafio aumenta em genômica, doenças raras, trajetórias muito específicas ou bases com grande granularidade temporal. Por isso, anonimização não deveria ser tratada como ritual administrativo que transforma qualquer conjunto em informação sem risco. É uma avaliação técnica e contextual: depende do que foi removido, do que permaneceu, de quais dados externos existem e de quais meios razoavelmente disponíveis poderiam ser usados para reidentificar indivíduos (PRICE; COHEN, 2019).

Há ainda uma camada menos intuitiva: o próprio modelo pode revelar informação sobre os dados com os quais aprendeu. Shokri et al. (2017) demonstraram os chamados ataques de inferência de participação, nos quais um adversário tenta estimar se determinado registro fez parte do conjunto de treinamento. O trabalho incluiu, entre os dados avaliados, informações de alta sensibilidade derivadas de altas hospitalares. A lição não é que todo modelo clínico necessariamente exponha seus pacientes, mas que treinamento não é uma transformação de mão única na qual os dados desaparecem depois de convertidos em parâmetros. Dependendo da arquitetura, do grau de sobreajuste, das interfaces disponibilizadas e das proteções adotadas, informações sobre o conjunto de treinamento podem permanecer parcialmente inferíveis.

Esse risco se torna ainda mais importante com modelos generativos, que produzem saídas extensas e podem estar conectados a documentos, bancos internos e ferramentas externas. A orientação da Organização Mundial da Saúde para grandes modelos multimodais inclui privacidade e cibersegurança entre os riscos que precisam ser considerados ao longo do ciclo de vida. Na prática, isso significa que inserir dados clínicos em uma ferramenta externa não pode ser tratado como equivalente a digitá-los em qualquer software institucional. É necessário saber que dados são enviados, onde são processados, se ficam registrados, para que podem ser reutilizados, quem possui acesso e quais obrigações contratuais e técnicas existem para protegê-los (OMS, 2024).

Tecnologias de preservação de privacidade procuram reduzir essa dependência de reunir todos os dados em um único lugar. A aprendizagem federada é um dos exemplos mais conhecidos: diferentes instituições podem treinar partes de um modelo mantendo os registros localmente e compartilhando atualizações do processo de aprendizagem, em vez de transferir diretamente toda a base. Em saúde, essa arquitetura é atraente justamente porque grandes conjuntos costumam estar distribuídos entre hospitais e sistemas que não podem simplesmente centralizar informações sensíveis. Rieke et al. (2020) descrevem o potencial dessa abordagem para a saúde digital, enquanto Kaissis et al. (2020) mostram que técnicas federadas precisam ser combinadas a medidas de segurança e proteção adequadas.

O detalhe é importante: aprendizagem federada não significa privacidade automática. Atualizações de modelo também podem carregar informação, participantes maliciosos podem tentar alterar o processo de treinamento e a infraestrutura de coordenação pode se tornar ponto de ataque. Dependendo da aplicação, mecanismos como agregação segura, controle de acesso, criptografia e privacidade diferencial podem ser necessários. O princípio mais útil é menos tecnológico do que parece: reduzir circulação desnecessária de dados é valioso, mas nenhuma arquitetura dispensa análise de risco. A pergunta correta não é se uma solução "usa privacidade", e sim quais ameaças ela reduz, quais permanecem e como serão monitoradas (KAISSIS et al., 2020).

No Brasil, esse cuidado se conecta diretamente à transformação digital do SUS. O Decreto nº 12.560/2025 regulamentou a Rede Nacional de Dados em Saúde e as Plataformas SUS Digital dentro de uma estrutura que envolve interoperabilidade, compartilhamento, proteção e governança de dados. O movimento é relevante porque mostra que integração e proteção não são objetivos mutuamente excludentes. Um

sistema de saúde fragmentado perde continuidade de cuidado; um sistema integrado sem governança amplia consequências de acessos indevidos e usos incompatíveis com a finalidade. A qualidade da infraestrutura digital depende de conseguir sustentar as duas exigências ao mesmo tempo (BRASIL, 2025).

Esse equilíbrio também muda a ideia de consentimento. Em saúde, nem todo tratamento de dado sensível depende de consentimento, porque a própria legislação prevê outras bases legítimas, inclusive em situações relacionadas à tutela da saúde e a políticas públicas. Reduzir privacidade a uma caixa de seleção transfere para o paciente uma responsabilidade que pertence à organização. Mesmo quando o consentimento é pertinente, ele precisa corresponder a uma escolha inteligível, e não a uma autorização genérica para usos futuros impossíveis de compreender. Proteção de dados exige finalidade, necessidade, segurança, transparência e governança; consentimento é uma ferramenta dentro desse conjunto, não um substituto para ele (BRASIL, 2018).

A pergunta sobre privacidade, portanto, conduz inevitavelmente a outra: como alguém pode avaliar riscos e exercer autonomia se não sabe onde a inteligência artificial está sendo usada, quais dados alimentam o sistema e qual papel sua saída ocupa no cuidado? É nesse ponto que transparência e explicabilidade entram na narrativa. Elas são frequentemente tratadas como sinônimos, mas respondem a problemas diferentes.

Transparência não é abrir a caixa-preta inteira

Transparência é uma relação entre informação e destinatário. Um pesquisador que avalia um modelo precisa conhecer detalhes que seriam inúteis para um paciente; um profissional precisa entender limites operacionais que talvez não exijam acesso ao código-fonte; uma instituição precisa saber versão, fornecedor, finalidade pretendida, requisitos de infraestrutura, histórico de atualizações e condições de monitoramento. Em 2024, FDA, Health Canada e MHRA formalizaram princípios de transparência para dispositivos médicos habilitados por aprendizado de máquina e definiram transparência como a comunicação clara de informações relevantes sobre finalidade, desenvolvimento, desempenho e, quando disponível, lógica do sistema. A explicabilidade aparece como uma dimensão possível dessa transparência, não como seu equivalente (FDA; HEALTH CANADA; MHRA, 2024).

Essa distinção evita uma armadilha comum. Quando um modelo é descrito como "caixa-preta", surge a ideia de que confiança dependeria de revelar cada operação interna. Na clínica, porém, compreender matematicamente cada parâmetro de uma

rede neural não é o que permite decidir se ela deve ser usada. O profissional precisa saber para qual população o sistema foi validado, que tarefa realiza, com qual comparador, quais erros são mais frequentes, como expressa incerteza, quando não deve ser usado e o que mudou desde a última versão. Em muitas situações, esse conjunto de informações é mais útil para segurança do que uma representação visual sofisticada do funcionamento interno.

Explicabilidade tenta responder a uma pergunta mais específica: por que determinado resultado foi produzido, ou quais elementos contribuíram para ele? Mapas de atenção, importância de variáveis e explicações locais podem ser úteis para investigação, auditoria ou interação. O problema começa quando uma explicação aproximada é tratada como reconstrução fiel do raciocínio do modelo. Ghassemi, Oakden-Rayner e Beam (2021) argumentam que abordagens atuais de explicabilidade podem criar uma falsa sensação de segurança em aplicações clínicas, porque explicações plausíveis não garantem que o sistema aprendeu uma relação clinicamente válida nem que seu comportamento será estável fora daquele caso.

Essa cautela conversa com temas já discutidos no livro. No Capítulo 3, vimos que uma variável altamente preditiva não precisa ser causal; no Capítulo 6, que desempenho retrospectivo não substitui validação; no Capítulo 7, que explicações podem modificar a maneira como usuários se ancoram em uma recomendação. O mesmo raciocínio vale aqui. Mostrar que creatinina, idade ou determinada região da imagem contribuíram para uma previsão não demonstra que alterar essas características modificará o desfecho. Uma explicação pode ajudar a inspecionar o sistema e ainda ser insuficiente para justificar uma conduta.

Amann et al. (2020) propõem que explicabilidade em saúde seja tratada como questão multidisciplinar porque diferentes atores buscam finalidades distintas: confiança, compreensão, responsabilização, aprendizado, contestação ou detecção de falhas. Isso muda o desenho desejável. Uma explicação para um radiologista pode destacar sinais técnicos que ajudem a revisar uma imagem; para um paciente, pode ser mais importante saber que a tecnologia foi usada, qual foi seu papel, quais limitações existem e quem permanece responsável pela decisão. Não existe uma única forma de explicação ideal para todos.

A regulamentação profissional brasileira recente segue direção semelhante. A Resolução CFM nº 2.454/2026, publicada em fevereiro de 2026 e com vigência prevista para 180 dias após a publicação, estabelece deveres de transparência e prevê

que médicos tenham acesso a informações claras sobre funcionamento, finalidade, limitações, riscos e grau de evidência dos sistemas utilizados. Também determina que o uso de IA seja comunicado e explicado aos pacientes em termos compreensíveis quando integrado ao cuidado. A norma não exige que todo paciente receba uma aula de ciência de dados; exige que informação relevante não desapareça justamente quando a tecnologia começa a influenciar uma decisão (CFM, 2026).

Transparência útil também precisa incluir o que muda ao longo do tempo. Um sistema identificado apenas pelo nome comercial pode receber atualização de modelo, alteração de limiar, nova base de conhecimento ou mudança de integração sem que o usuário perceba. Se o desempenho depende de versão e contexto, rastreabilidade precisa registrar qual sistema produziu determinada saída e em que condições. Isso é especialmente importante em modelos generativos, nos quais pequenas mudanças de instrução, contexto ou versão podem modificar respostas. Sem rastreabilidade, investigar um incidente depois que ele ocorreu se torna muito mais difícil.

Há, contudo, um limite legítimo. Transparência não exige exposição indiscriminada de segredos comerciais, código, credenciais ou detalhes que criem novas vulnerabilidades. O próprio princípio regulatório precisa equilibrar informação suficiente para uso seguro com proteção de propriedade intelectual e de segurança. A questão central é se o segredo impede o usuário de conhecer algo material para o risco clínico. Uma instituição pode não conhecer cada peso de uma rede neural e ainda assim exigir documentação de validação, composição dos dados, limites de uso, histórico de mudanças, controles de acesso, plano de monitoramento e procedimento de resposta a incidentes.

Esse último item aproxima transparência de cibersegurança. Um sistema pode ser validado, bem documentado e clinicamente útil e ainda se tornar inseguro se credenciais forem comprometidas, dados forem alterados, componentes deixarem de receber correções ou um ataque modificar sua entrada. Na medicina digital, segurança da informação não é um problema paralelo ao cuidado. Ela participa diretamente da segurança do paciente.

Quando segurança digital vira segurança clínica

A maneira mais simples de perceber essa ligação é pensar em três propriedades clássicas da segurança da informação: confidencialidade, integridade e disponibilidade. Confidencialidade responde a quem consegue acessar uma informação; integridade, a se dados e sistemas permanecem corretos e não foram alterados de forma indevida;

disponibilidade, a se continuam acessíveis quando o cuidado precisa deles. Em saúde, as três podem ter consequência clínica. Um vazamento expõe intimidade e pode produzir discriminação; um dado alterado pode induzir uma conduta errada; um sistema indisponível pode atrasar diagnóstico, medicação ou monitoramento.

A Organização Mundial da Saúde, em relatório de 2024 sobre ameaças futuras de cibersegurança em saúde digital, chama atenção para o fato de que digitalização crescente amplia dependência de infraestrutura conectada e exige preparação capaz de preservar continuidade dos serviços. Essa perspectiva é importante porque cibersegurança costuma ser lembrada apenas depois de um vazamento. Em um hospital, entretanto, o incidente mais perigoso nem sempre é aquele que torna dados públicos. Pode ser aquele que impede acesso ao prontuário, paralisa um equipamento, corrompe uma informação ou faz uma saída manipulada parecer legítima (WHO, 2024).

Sistemas de inteligência artificial acrescentam vetores próprios a riscos já conhecidos de software. O NIST organiza ameaças de aprendizado de máquina adversarial ao longo do ciclo de vida e inclui manipulação de dados de treinamento, entradas construídas para induzir comportamento incorreto, ataques contra a privacidade do treinamento e manipulações de sistemas generativos conectados a dados ou ferramentas. Essas categorias ajudam porque impedem que "ataque à IA" seja tratado como fenômeno único. Um envenenamento de dados ameaça a integridade do que o modelo aprende; uma inferência de participação ameaça confidencialidade; uma manipulação de entrada pode afetar segurança do resultado; indisponibilidade atinge o próprio serviço (NIST, 2025).

Nem todo risco adversarial exige um atacante altamente sofisticado. Erros de configuração, permissões excessivas, bibliotecas desatualizadas, credenciais compartilhadas e integrações mal controladas podem ser tão relevantes quanto técnicas específicas contra modelos. A segurança de uma aplicação de IA é a segurança do sistema inteiro que a sustenta: banco de dados, rede, serviço em nuvem, interface, dispositivos conectados, sistemas de autenticação, componentes de terceiros e processos humanos. Avaliar apenas o algoritmo seria equivalente a avaliar a segurança de uma unidade de terapia intensiva examinando somente um monitor.

A expansão de dispositivos conectados torna essa visão de sistema ainda mais concreta. A orientação final da FDA sobre cibersegurança em dispositivos médicos, atualizada em 2026, trata segurança como responsabilidade de ciclo de vida: desenho

seguro, documentação de riscos, gestão de vulnerabilidades, atualizações e correções pós-mercado. Para determinados "cyber devices", a legislação norte-americana exige ainda informações como uma lista de materiais de software, que ajuda a identificar componentes comerciais, de código aberto ou de terceiros presentes no dispositivo. A lógica é simples: uma vulnerabilidade em um componente invisível para a instituição continua sendo uma vulnerabilidade do produto (FDA, 2026c).

Essa abordagem de ciclo de vida combina com o que o Capítulo 6 mostrou sobre drift. Um modelo não se torna permanentemente seguro no dia em que é instalado; tampouco um dispositivo se torna permanentemente ciberseguro no dia em que passa por avaliação. Novas vulnerabilidades são descobertas, bibliotecas deixam de receber suporte, integrações mudam e atacantes aprendem. Monitorar desempenho sem monitorar segurança deixaria aberta uma via pela qual o comportamento do sistema pode mudar sem alteração aparente do modelo.

Em modelos generativos, o risco ganha outra forma porque a linguagem se torna também uma interface de controle. Um sistema pode receber textos externos, pesquisar bases, resumir documentos e, em arquiteturas mais avançadas, acionar ferramentas. Instruções maliciosas incorporadas ao conteúdo processado podem tentar redirecionar seu comportamento; dados sensíveis presentes no contexto podem ser expostos a saídas indevidas; permissões excessivas podem transformar uma resposta errada em ação. O relatório do NIST sobre aprendizado adversarial diferencia justamente modelos generativos porque sua integração a dados privados e ferramentas cria combinações de risco que não existiam da mesma maneira em classificadores tradicionais (NIST, 2025).

A resposta institucional precisa ser proporcional à consequência. Criptografia, autenticação forte, princípio do menor privilégio, segregação de ambientes, registro de atividades, gestão de fornecedores, atualização de componentes, cópias de segurança e testes de recuperação são controles conhecidos, mas seu valor depende de como se relacionam ao cuidado. Um backup que existe mas não pode ser restaurado no tempo necessário para a assistência oferece uma proteção teórica. Um log que registra acessos mas nunca é revisado não produz vigilância. Uma política que proíbe determinadas ferramentas sem oferecer alternativa segura pode simplesmente deslocar o uso para fora da visibilidade institucional.

Também é necessário planejar o momento em que a proteção falha. Incidentes precisam ter critérios de detecção, responsáveis, canais de comunicação, estratégias de

contenção, continuidade assistencial e processo de aprendizagem posterior. No Brasil, a Resolução CD/ANPD nº 15/2024 regulamenta a comunicação de incidentes de segurança que possam acarretar risco ou dano relevante aos titulares. No campo médico, a Resolução CFM nº 2.454/2026 prevê proteção de dados, modelos, sistemas e ambientes computacionais contra destruição, perda, alteração, acesso não autorizado e vazamentos. As duas perspectivas convergem em um ponto: segurança não termina na prevenção; inclui capacidade de responder quando o evento ocorre (ANPD, 2024; CFM, 2026).

Essa capacidade depende de responsabilidades previamente definidas. Se uma ferramenta clínica utiliza infraestrutura de um hospital, modelo de um fornecedor, serviço de nuvem de terceiro e integração desenvolvida por outra empresa, um incidente pode atravessar várias fronteiras contratuais. A equipe assistencial precisa saber a quem recorrer; a instituição precisa conhecer dependências técnicas; fornecedores precisam ter obrigações de notificação e correção; e pacientes não podem ser deixados diante de uma cadeia em que todos apontam para outro ator. Cibersegurança é, nesse sentido, um problema de governança antes de ser apenas um problema de engenharia.

Privacidade, transparência, explicabilidade e cibersegurança são, portanto, conceitos diferentes, mas interdependentes. Privacidade pergunta se o uso do dado respeita finalidade, direitos e limites; transparência pergunta se informações necessárias chegam a quem precisa delas; explicabilidade tenta tornar aspectos do comportamento do sistema compreensíveis; cibersegurança protege dados, modelos e infraestrutura contra acesso, alteração e indisponibilidade indevidos. Uma dimensão pode fortalecer outra, mas não substituí-la. Um modelo explicável pode ser inseguro. Um sistema criptografado pode ser opaco. Um processo transparente pode coletar dados além do necessário. Um banco anonimizado pode produzir um modelo clinicamente inútil se a proteção destruir justamente os sinais relevantes.

A excelência, portanto, não está em maximizar isoladamente cada propriedade. Está em construir uma arquitetura em que proteção não impeça cuidado necessário, transparência não se transforme em sobrecarga informacional, explicação não seja usada como maquiagem para validação insuficiente e segurança não seja lembrada apenas depois do incidente. Esse equilíbrio precisa ser definido antes da implantação e revisto quando dados, modelos, fornecedores e usos mudam. É uma tarefa contínua, tão sociotécnica quanto as questões de validação e equidade discutidas nos capítulos anteriores.

Essa arquitetura torna inevitável uma pergunta de responsabilidade. Se um modelo é desenvolvido por uma empresa, treinado com dados fornecidos por instituições, integrado ao prontuário por outra equipe, configurado localmente, utilizado por profissionais e atualizado ao longo do tempo, quem responde quando a recomendação causa dano, quando a informação é insuficiente ou quando a governança falha? Construir sistemas confiáveis, nesse cenário, também significa distribuir de modo explícito deveres, autoridade e capacidade de intervenção entre os diferentes atores que participam do cuidado.

CAPÍTULO 9
QUEM RESPONDE PELA DECISÃO? REGULAÇÃO, RESPONSABILIDADE PROFISSIONAL E GOVERNANÇA DA IA EM SAÚDE

Responsabilidade não cabe em uma única pessoa

Essa pergunta não admite uma resposta simples porque responsabilidade e proximidade com o paciente não coincidem necessariamente. Em uma solução de IA, decisões sobre treinamento, hospedagem, integração, configuração local e uso profissional podem ser tomadas por atores diferentes e em momentos distintos. Quando tudo funciona, essa distribuição de tarefas parece eficiente. Quando ocorre dano, porém, surge a questão decisiva: em qual ponto da cadeia termina a contribuição técnica e começa a responsabilidade por aquilo que aconteceu?

A resposta não pode ser obtida simplesmente escolhendo a pessoa que estava mais próxima do paciente. Em sistemas sociotécnicos, um resultado clínico pode depender de decisões tomadas muito antes da consulta: qual população foi usada no treinamento, qual desfecho foi definido como alvo, que limiar transformou uma probabilidade em alerta, quais atualizações foram instaladas, como a interface destacou uma recomendação e que informação o hospital ofereceu ao usuário sobre limites conhecidos. O profissional continua responsável por sua atuação, mas não controla sozinho todas essas camadas. Tratar qualquer evento adverso como “erro médico com IA” pode esconder falhas de projeto, implantação ou governança que nenhum julgamento individual seria capaz de corrigir.

A literatura mostra que a própria palavra responsabilidade é usada de maneiras diferentes. Bouhouita-Guermech e Haidar (2024), em uma revisão de escopo com 136 estudos sobre IA em saúde, encontraram sobreposição entre responsabilidade moral, profissional, causal, organizacional e jurídica. Cestonaro et al. (2023), ao revisar a discussão médico-legal sobre algoritmos diagnósticos, chegaram a uma conclusão semelhante: não existe resposta única e consolidada para a distribuição de responsabilidade quando dano resulta da interação entre profissional e sistema. Essa incerteza não significa ausência de deveres. Significa que deveres diferentes recaem sobre atores diferentes e precisam ser identificados antes que um incidente obrigue todos a reconstruí-los retrospectivamente.

Há uma diferença importante entre responder por uma decisão clínica e responder pela confiabilidade do sistema que participou dela. O médico pode ter o dever de confrontar uma recomendação com história, exame, evidência e contexto; a instituição, por sua vez, escolhe quais sistemas estarão disponíveis, define fluxos, treinamento, permissões e mecanismos de escalonamento; o desenvolvedor decide arquitetura, dados, testes, documentação e políticas de atualização; o fornecedor controla versões, manutenção e suporte; reguladores estabelecem requisitos para determinadas categorias de produto. Essas responsabilidades podem se encontrar no mesmo evento sem se tornarem idênticas.

Essa distinção é necessária para evitar dois extremos. O primeiro é imaginar que a presença de inteligência artificial transfere a decisão para a máquina e, com ela, a responsabilidade. Um sistema não assume dever ético perante o paciente, não responde a processo disciplinar, não indeniza dano e não participa de uma conversa sobre valores. O segundo extremo é declarar que, porque o humano mantém a palavra final, todo risco tecnológico passa automaticamente a ser dele. Fotheringham e Smith (2024) chamam atenção, no contexto britânico, para o risco de profissionais se tornarem pontos de absorção de responsabilidade por falhas que surgem em sistemas sobre os quais possuem controle limitado. Embora a solução jurídica dependa de cada país, o problema organizacional é mais amplo: supervisão humana não pode servir como mecanismo para transferir ao usuário final riscos que deveriam ter sido controlados em outras etapas.

A Resolução CFM nº 2.454/2026 é especialmente interessante porque torna essa tensão explícita no contexto brasileiro. Publicada em 27 de fevereiro de 2026, a norma ainda estava em vacatio legis em 22 de agosto do mesmo ano, pois seu art. 23 estabelece vigência após 180 dias da publicação. O texto afirma que o médico deve utilizar a IA como instrumento de apoio, manter julgamento crítico e permanecer responsável pelos atos médicos praticados com auxílio desses sistemas. Ao mesmo tempo, reconhece o direito de recusar ferramentas sem validação ou certificação pertinente e prevê proteção contra responsabilização indevida por falhas atribuíveis exclusivamente ao sistema quando houver uso diligente, crítico e ético. A exposição de motivos também admite, conforme o caso, responsabilização de desenvolvedores ou fornecedores. A leitura correta, portanto, é mais sofisticada do que "o médico responde por tudo" (CFM, 2026).

A própria resolução desloca parte do foco para as instituições. Ela prevê avaliação preliminar de risco, processos internos de governança, auditoria, monitoramento e

supervisão humana. Para instituições que adotarem sistemas próprios de IA, determina a criação de Comissão de IA e Telemedicina sob coordenação médica e subordinada à direção técnica. O anexo dedicado à governança atribui ao diretor técnico responsabilidade pelas diretrizes de segurança, ética e transparência e inclui monitoramento de vieses, gestão do ciclo de vida, documentação e acesso a informações por órgãos de controle. Mesmo antes de sua entrada em vigor, essa arquitetura revela uma mudança importante: segurança não é apresentada como virtude individual do médico que "usa bem" uma ferramenta, mas como propriedade que precisa ser construída institucionalmente (CFM, 2026).

Isso também muda o significado de supervisão humana. Um profissional só pode revisar de maneira substancial uma recomendação se tiver tempo, informação e possibilidade real de discordar. Se a interface mostra apenas uma conclusão sem indicar finalidade, incerteza ou limites; se a organização cria metas que favorecem adesão automática; se o sistema é utilizado em volume incompatível com revisão cuidadosa; ou se a contestação exige um caminho impraticável, a presença formal de um humano não garante controle. A OMS já havia chamado atenção para a necessidade de pontos reais de supervisão e de responsabilização ao longo do ciclo de vida, e não apenas para a assinatura final de um profissional (OMS, 2021).

O registro do uso também faz parte dessa responsabilidade. Se uma recomendação de IA influenciou uma decisão relevante, reconstruir o evento posteriormente exige saber qual sistema foi usado, em qual versão, com quais dados de entrada, qual saída foi produzida e como ela foi interpretada. Sem rastreabilidade, um incidente se torna uma disputa de memória. O profissional pode lembrar que "o sistema sugeriu", o fornecedor pode alegar que a versão atual funciona de outra forma e a instituição pode não conseguir demonstrar quais configurações estavam ativas naquele momento. A governança começa a se aproximar, então, da própria capacidade de provar o que aconteceu.

Essa perspectiva ajuda a separar responsabilidade de culpa. O objetivo de registrar versões, decisões e intervenções não deveria ser construir antecipadamente um culpado, mas tornar possível aprender com falhas e reparar danos quando necessário. Em sistemas complexos, responsabilização madura não é a arte de encontrar uma pessoa para carregar todo o evento; é a capacidade de identificar quais deveres existiam, quem possuía condições de cumpri-los, onde a barreira de segurança falhou e que mudança reduz a probabilidade de repetição.

Regular o risco sem congelar a medicina

Se a responsabilidade é distribuída, a regulação também tende a ser. Não existe uma única "lei da inteligência artificial em saúde" capaz de substituir normas sanitárias, profissionais, de proteção de dados, pesquisa, segurança da informação e direitos do paciente. O mesmo sistema pode estar simultaneamente submetido a exigências diferentes conforme sua finalidade. Um software que agenda consultas, um modelo que resume prontuários e uma ferramenta que determina probabilidade de câncer podem utilizar técnicas semelhantes de aprendizado de máquina, mas não produzem o mesmo risco e não deveriam receber o mesmo tratamento regulatório.

A regulação sanitária brasileira já trabalha com essa lógica de finalidade. A RDC nº 657/2022 da Anvisa disciplina software como dispositivo médico e não transforma todo programa que utiliza IA em dispositivo médico. O enquadramento depende da finalidade pretendida e da função médica desempenhada. Essa distinção é decisiva porque impede que "ter inteligência artificial" se torne o critério principal. O que importa é o que o produto faz, para quem, com que consequência e em qual posição do cuidado. Um algoritmo que gera material administrativo não deve ser avaliado como um sistema que orienta tratamento apenas porque ambos utilizam redes neurais (ANVISA, 2022a; ANVISA, 2022b).

O cenário brasileiro geral permanece em construção. Em 22 de agosto de 2026, o Projeto de Lei nº 2.338/2023, aprovado pelo Senado, continuava em tramitação na Câmara dos Deputados, sob análise de comissão especial, e não constituía uma lei geral de inteligência artificial em vigor. Essa precisão temporal importa. Em um campo que muda rapidamente, apresentar projeto legislativo como norma vigente produz uma falsa sensação de estabilidade jurídica e pode tornar um livro desatualizado já no momento de sua publicação. O mais correto é reconhecer que a governança brasileira combina, até aqui, regras setoriais já existentes com uma legislação geral ainda em debate (BRASIL, 2026).

A União Europeia oferece um exemplo diferente: criou um regulamento horizontal baseado em risco, mas sua aplicação é escalonada e dialoga com legislações setoriais. O Regulamento (UE) 2024/1689 entrou em vigor em 2024 e passou a aplicar diferentes conjuntos de obrigações em etapas. A versão consolidada vigente em agosto de 2026 já refletia alterações recentes no cronograma. As regras para determinados sistemas de alto risco do Anexo III foram deslocadas para dezembro de 2027, enquanto as relativas a sistemas de alto risco incorporados a produtos sujeitos à

legislação harmonizada, categoria que inclui dispositivos médicos em determinadas condições, passaram a ter aplicação posterior, em agosto de 2028. O detalhe não é meramente burocrático: mostra como mesmo um marco regulatório abrangente precisa ser ajustado à disponibilidade de padrões, estruturas de conformidade e legislação de produto (UNIÃO EUROPEIA, 2024).

Nos Estados Unidos, a FDA segue outra arquitetura. Em vez de depender de uma lei geral de IA para cada aplicação médica, utiliza competências já existentes sobre dispositivos e software e vem adaptando uma abordagem de ciclo de vida. Os planos predeterminados de controle de mudanças permitem que certas modificações previstas em funções de software de dispositivos habilitados por IA sejam descritas, validadas e controladas antecipadamente. A lógica reconhece que modelos podem evoluir, mas rejeita a ideia de que evolução significa liberdade para mudar sem rastreabilidade ou supervisão. A flexibilidade regulatória aparece, assim, associada a uma obrigação maior de planejar como a mudança será produzida e demonstrada (FDA, 2025).

Esses modelos regulatórios não precisam ser tratados como competidores em busca de uma única fórmula correta. Eles respondem a tradições jurídicas e institucionais diferentes, mas convergem em alguns princípios: finalidade pretendida, proporcionalidade ao risco, documentação, validação, supervisão, transparência, segurança e acompanhamento durante o ciclo de vida. A convergência é mais relevante para a clínica do que a comparação abstrata entre jurisdições. Um hospital que pretende incorporar um sistema precisa transformar esses princípios em perguntas operacionais independentemente do país: o produto está autorizado para esta finalidade? A população local se aproxima daquela em que foi validado? Quem aprova atualizações? Que evento obriga suspensão? Como será comunicada uma falha?

A autorização regulatória, contudo, não resolve a decisão de adoção. Um dispositivo pode atender aos requisitos para ser colocado no mercado e ainda não ser a melhor opção para determinado serviço. O Capítulo 6 mostrou que desempenho depende de prevalência, população, equipamento e fluxo. Regulação responde se uma tecnologia pode ser disponibilizada sob determinadas condições; governança institucional precisa responder se ela deve ser usada naquele ambiente e como. Confundir as duas etapas cria uma forma de terceirização: “se foi aprovado, deve ser seguro para nós”. A aprovação é evidência importante, não substituto para avaliação local.

O contrário também é verdadeiro. Um sistema que não se enquadra como dispositivo médico não se torna automaticamente isento de risco. Modelos generativos usados para redigir mensagens, resumir prontuários ou apoiar raciocínio podem produzir dano sem emitir formalmente um diagnóstico. A classificação regulatória de produto e o risco assistencial não são mapas perfeitamente sobrepostos. Uma ferramenta administrativa pode expor dados sensíveis; um assistente de documentação pode inserir uma informação falsa que depois influenciará condutas; um modelo de finalidade geral pode ser levado, pelo próprio usuário, a responder perguntas clínicas para as quais nunca foi validado. Governança precisa enxergar essas zonas antes que elas sejam formalmente enquadradas por uma categoria de dispositivo.

Essa é uma das razões pelas quais regras baseadas em risco são atraentes. Elas permitem concentrar exigências onde consequência, autonomia, escala e dificuldade de reversão são maiores. Mas classificação de risco também pode falhar se for tratada como etiqueta permanente. Uma ferramenta de baixo impacto pode mudar de função quando passa a integrar o prontuário; um modelo usado apenas para apoio educacional pode começar a orientar decisões reais; uma aplicação experimental pode se tornar rotina sem que ninguém tenha registrado a transição. O risco não está apenas no produto. Está no uso concreto que a organização faz dele.

Regular bem, portanto, não significa produzir o maior número possível de proibições. Significa construir fronteiras claras para inovação responsável. Regras muito vagas podem deixar pacientes e profissionais sem proteção; regras impossíveis de cumprir podem empurrar usos para ambientes informais, fora de qualquer supervisão. O desafio é permitir experimentação controlada, exigir evidência proporcional à consequência e impedir que a velocidade comercial seja confundida com maturidade clínica. Essa lógica aproxima regulação de governança: ambas tentam definir, antes do dano, quais condições tornam um uso aceitável.

Governança é decidir antes do incidente

Governança costuma aparecer em documentos como uma palavra ampla, quase sinônimo de “fazer tudo corretamente”. Para ser útil, ela precisa se tornar concreta. Em uma instituição de saúde, governar IA começa por saber que sistemas existem. Parece trivial, mas ferramentas podem entrar por diferentes portas: contratação institucional, módulo acrescentado a um prontuário, recurso ativado por fornecedor, plataforma de pesquisa, extensão de navegador, conta pessoal de um profissional ou

modelo generativo usado informalmente para acelerar tarefas. Se a organização não consegue construir um inventário razoável, não consegue classificar risco, acompanhar versões nem saber onde um incidente pode ter ocorrido.

O inventário, porém, é apenas o começo. Cada sistema precisa ter uma finalidade autorizada, um responsável institucional, documentação de validação, regras de acesso, população de uso, fluxos de revisão, métricas de monitoramento e critérios para suspensão ou retirada. Uma mudança de versão precisa ter dono. Um alerta de segurança precisa chegar a alguém. Uma deterioração de desempenho precisa produzir ação previamente combinada. Governança é justamente transformar frases como "o sistema será monitorado" em uma cadeia na qual se sabe quem monitora, o que observa, com qual frequência, que limiar exige investigação e quem possui autoridade para interromper o uso.

Esse trabalho começa antes da compra. Contratar uma solução de IA não é adquirir apenas uma licença; é importar uma parte da governança do fornecedor para dentro do cuidado. O contrato precisa deixar visível que dados serão processados, se poderão ser reutilizados, como atualizações serão comunicadas, quais evidências serão fornecidas, em quanto tempo incidentes relevantes serão notificados, que apoio existirá para investigação e como a instituição recuperará seus dados e registros se encerrar a relação. Um sistema tecnicamente excelente pode se tornar operacionalmente inseguro quando a organização não consegue saber que versão está rodando ou depende de um fornecedor que não oferece rastreabilidade suficiente.

A preocupação é ainda maior com modelos generativos e ferramentas de fácil acesso. Comeau, Bitterman e Celi (2025) alertam para o risco de uma espécie de experimentação clínica difusa quando sistemas são incorporados a fluxos assistenciais sem transparência, monitoramento e estruturas formais de responsabilização. O problema não exige que alguém declare estar "fazendo pesquisa". Se centenas de profissionais começam a usar uma ferramenta não avaliada para resumir prontuários, orientar respostas ou sugerir hipóteses, a instituição pode estar alterando o cuidado de forma sistemática sem ter definido desfechos, critérios de segurança ou mecanismo para aprender com erros.

Proibir de forma absoluta costuma ser uma resposta frágil. Ferramentas úteis tendem a reaparecer em contas pessoais, dispositivos externos ou fluxos impossíveis de auditar. A alternativa mais madura é oferecer caminhos seguros: soluções aprovadas para finalidades definidas, ambientes protegidos, treinamento, regras claras para dados

sensíveis e canais para solicitar avaliação de novas aplicações. Governança precisa reduzir o incentivo ao uso invisível. Quando a política institucional é apenas "não use", sem alternativa compatível com a necessidade real de trabalho, a organização pode preservar uma aparência de controle enquanto perde visibilidade sobre o que de fato acontece.

Comissões de IA podem ajudar, mas não devem se tornar instâncias cerimoniais. A previsão do CFM de uma Comissão de IA e Telemedicina em instituições com sistemas próprios aponta para a necessidade de coordenação, mas a qualidade dessa estrutura dependerá de composição e autoridade. Segurança clínica, tecnologia da informação, proteção de dados, engenharia clínica, qualidade, jurídico, pesquisa, gestão e assistência enxergam riscos diferentes. Em sistemas de impacto relevante, também faz sentido considerar perspectivas de pacientes e usuários. A governança se enfraquece quando a decisão fica restrita ao grupo que desenvolveu ou comprou a tecnologia e, por isso, tende naturalmente a enxergar mais benefícios do que modos de falha.

A autoridade para interromper também precisa ser real. Se um modelo apresenta piora relevante em um subgrupo, começa a gerar saídas inconsistentes após atualização ou participa de incidente grave, alguém deve poder colocá-lo em modo de revisão sem depender de uma negociação improvisada com quem financiou o projeto. Isso pode gerar custos e frustração, mas é justamente a diferença entre monitoramento e observação passiva. Um painel que mostra deterioração sem mecanismo de resposta é apenas uma forma elegante de assistir ao problema acontecer.

Governança também precisa proteger o profissional que discorda da automação. A Resolução CFM nº 2.454/2026 afirma que o médico pode acolher ou rejeitar sugestões da IA conforme seu julgamento e veda políticas institucionais que subordinem condutas de forma automática. Esse ponto é mais profundo do que parece. Se discordar do algoritmo reduz produtividade, exige justificativas excessivas ou expõe o profissional a punição informal, a autonomia existe no papel e desaparece no fluxo. A instituição precisa observar não apenas se há um botão para rejeitar a recomendação, mas se o ambiente realmente permite utilizá-lo (CFM, 2026).

O mesmo vale para o paciente. Informar que "foi utilizada inteligência artificial" é insuficiente se a informação não esclarece o papel relevante da ferramenta. Em algumas situações, bastará explicar que um sistema ajudou a organizar documentação; em outras, será importante dizer que uma estimativa algorítmica contribuiu para

priorização ou escolha clínica. O grau de informação precisa acompanhar o impacto. A OMS insiste que autonomia e responsabilidade exigem inteligibilidade apropriada para cada destinatário, e a norma do CFM prevê informação clara quando a IA tiver papel relevante no cuidado. Transparência, aqui, não é produzir um relatório técnico para o paciente; é permitir que ele compreenda o suficiente para participar, questionar e buscar segunda opinião quando necessário (OMS, 2021; CFM, 2026).

Quando ocorre um incidente, uma governança bem desenhada muda a qualidade da investigação. Em vez de perguntar apenas "quem clicou?", a organização pode reconstruir dados de entrada, versão, configuração, saída, decisão humana, atualização recente, contexto de trabalho e evidências disponíveis naquele momento. Essa reconstrução torna possível distinguir defeito do produto, problema de integração, uso fora da finalidade, falha de treinamento, viés de automação ou combinação entre fatores. Ela também permite que aprendizado ultrapasse o caso individual. Se a causa está na interface ou no fluxo, treinar novamente um único profissional não resolve.

Responsabilidade distribuída não pode significar responsabilidade diluída. Quanto mais atores participam do sistema, mais importante se torna definir deveres explícitos. O desenvolvedor não deve poder apontar para o hospital quando o problema é uma alteração não comunicada; a instituição não deve apontar para o médico quando impôs uma ferramenta sem validação local; o profissional não deve usar a opacidade como justificativa para ignorar uma inconsistência clínica evidente; e o regulador não deve tratar autorização inicial como substituto de vigilância. Accountability só existe quando cada ator consegue responder por aquilo que efetivamente controla e quando há mecanismos para conectar essas respostas.

Esse desenho também protege a inovação. Ambiguidade permanente sobre responsabilidade pode produzir dois resultados ruins: adoção impulsiva, porque todos presumem que outro ator absorverá o risco, ou paralisia defensiva, porque ninguém aceita participar de uma tecnologia cujo risco jurídico e institucional é impossível de delimitar. Governança reduz essa incerteza ao tornar expectativas explícitas. Ela não elimina eventos adversos, mas transforma o uso da IA de experimento informal em processo no qual decisões, mudanças e falhas podem ser justificadas, auditadas e corrigidas.

Ao final deste percurso, a pergunta "quem responde pela decisão?" recebe uma resposta menos simples, mas mais fiel à realidade. O médico continua responsável por seu ato profissional; a instituição responde pelas condições em que escolhe, implanta e

monitora tecnologia; desenvolvedores e fornecedores respondem por aspectos do sistema que projetam, modificam e disponibilizam; autoridades regulatórias definem limites e mecanismos de controle; e pacientes precisam permanecer sujeitos da decisão, não meros destinatários de uma cadeia automatizada. Essas responsabilidades se sobrepõem sem desaparecer umas nas outras.

A boa governança, portanto, não existe para colocar uma pessoa entre o algoritmo e o dano e chamá-la de responsável. Existe para construir um sistema em que a influência da inteligência artificial seja proporcional à evidência, visível para quem precisa avaliá-la, contestável quando necessário e rastreável quando alguma coisa falha. Quando essa arquitetura funciona, a tecnologia pode aumentar capacidade sem apagar autoria e dever. Quando ela falha, a automação corre o risco de produzir decisões às quais todos contribuíram, mas pelas quais ninguém consegue responder adequadamente.

Quando a influência da inteligência artificial é distribuída dessa forma, colaboração humano-máquina deixa de ser apenas uma questão de desempenho. Passa a envolver julgamento, autonomia, vínculo e dever. Uma arquitetura responsável precisa ampliar capacidades sem empobrecer justamente os elementos que tornam o cuidado reconhecível como cuidado: possibilidade de discordar, compreender limites, responder por consequências e preservar o paciente como participante da decisão.

CAPÍTULO 10
CUIDADO AUMENTADO, NÃO SUBSTITUÍDO: COLABORAÇÃO HUMANO-IA E CAMINHOS PARA UMA SAÚDE CONFIÁVEL E SUSTENTÁVEL

A melhor parceria não é a mais obediente

O capítulo anterior terminou em uma arquitetura de responsabilidade: profissionais, instituições, desenvolvedores e reguladores precisam saber o que lhes cabe quando a inteligência artificial participa do cuidado. O passo final desta obra é perguntar o que acontece antes do incidente, durante a consulta, na leitura de um exame, na revisão de um alerta ou na escolha de uma conduta. É nesse espaço cotidiano que a expressão colaboração humano-IA deixa de ser uma ideia abstrata e se torna desenho de trabalho. Colaborar não significa dividir uma tarefa ao meio nem aceitar que a máquina faça a parte "técnica" enquanto a pessoa preserva uma suposta parte "humana". Significa organizar capacidades diferentes de modo que uma possa revelar limites da outra sem apagar quem responde pela decisão.

Essa formulação é importante porque boa parte da discussão pública ainda se organiza em torno de uma competição: algoritmo contra médico, máquina contra especialista, automação contra trabalho humano. Os capítulos anteriores mostraram por que essa comparação é frequentemente insuficiente. Em alguns cenários, como o rastreamento autônomo de retinopatia diabética, a tecnologia pode ocupar uma etapa sem leitura especialista obrigatória. Em outros, como mamografia, alertas prognósticos, modelos generativos e suporte terapêutico, o objeto relevante é o sistema formado por ferramenta, usuário, fluxo e organização. Topol (2019) já propunha a convergência entre inteligência humana e artificial como horizonte mais produtivo do que a substituição. O avanço recente da evidência permite tornar essa ideia menos retórica: a qualidade da parceria depende da forma como o trabalho é distribuído.

Os ensaios com grandes modelos de linguagem ilustram bem esse ponto. No estudo de Goh et al. (2024), simplesmente disponibilizar um modelo a médicos que resolviam casos diagnósticos complexos não produziu melhora estatisticamente significativa em relação aos recursos convencionais. Em tarefas de manejo avaliadas posteriormente, o mesmo grupo encontrou benefício com assistência do GPT-4,

embora acompanhado de maior tempo por caso (GOH et al., 2025). Qazi et al. (2026) acrescentaram uma variável que não pode ser tratada como detalhe: médicos que haviam concluído treinamento estruturado em alfabetização de IA aproveitaram melhor o apoio do modelo em uma avaliação randomizada. Juntos, esses estudos indicam que "ter acesso à IA" não é uma intervenção suficientemente definida. A tarefa, a preparação do usuário e o modo de interação mudam o resultado.

Em 2026, Everett et al. testaram essa hipótese de maneira ainda mais explícita. Setenta médicos participaram de um ensaio randomizado com um sistema desenhado para colaboração diagnóstica estruturada. Médico e IA produziam avaliações independentes; depois, o sistema apresentava uma síntese com concordâncias, divergências e críticas aos dois raciocínios. Em um fluxo, a IA oferecia a primeira opinião; no outro, o clínico respondia antes de ver a análise do modelo. As duas formas de colaboração superaram os recursos convencionais em acurácia diagnóstica, com 85% e 82% contra 75%, respectivamente, sem diferença estatisticamente significativa entre os dois fluxos e sem superar de forma demonstrada o desempenho do modelo isolado. O estudo utilizou casos simulados, não desfechos de pacientes reais, mas sua contribuição é conceitualmente valiosa: interação pode ser projetada para produzir comparação crítica, e não mera obediência.

Essa diferença muda o sentido de "humano no circuito". Um profissional formalmente presente pode exercer pouca supervisão real se recebe centenas de recomendações, dispõe de segundos para avaliá-las e trabalha em uma interface que apresenta a saída algorítmica como padrão a ser confirmado. Nessa situação, a presença humana existe no organograma, mas não necessariamente no raciocínio. O risco de automação indevida discutido por Goddard, Roudsari e Wyatt (2012) surge justamente quando a influência da tecnologia é maior do que a capacidade prática de contestá-la. A colaboração de alta qualidade, portanto, não deve medir sucesso pelo número de vezes em que humano e modelo concordam. Concordância pode refletir competência compartilhada, mas também ancoragem, dependência ou ausência de revisão.

O objetivo mais importante é confiança calibrada. Confiar demais em uma ferramenta errada e desconfiar sistematicamente de uma ferramenta útil são falhas diferentes, mas ambas desperdiçam informação. Nicolson et al. (2025), em um estudo com profissionais que estimavam idade gestacional por ultrassonografia, mostraram que a predição do modelo melhorou o desempenho médio dos participantes, enquanto a adição de explicações produziu respostas heterogêneas: alguns

melhoraram, outros pioraram, e o aumento de confiança subjetiva não correspondeu automaticamente a melhor confiança ou dependência apropriada. O achado ajuda a desfazer outra simplificação frequente. Explicar mais não garante colaborar melhor. Uma explicação só possui valor clínico quando melhora uma decisão, ajuda a identificar uma falha ou permite ao usuário ajustar a confiança de maneira apropriada.

Por isso, uma boa interface deveria facilitar desacordo informado. Em determinadas tarefas, pode ser útil preservar uma primeira avaliação independente antes de revelar a sugestão automatizada; em outras, pode fazer sentido mostrar tendência temporal, incerteza, comparadores relevantes ou condições que colocam o paciente fora do domínio validado. O formato depende da tarefa. O princípio é mais estável: a tecnologia deve aumentar a capacidade de perceber evidências e contradições, não reduzir a decisão a uma confirmação rápida. O estudo de Everett et al. (2026) é interessante justamente porque incorpora a divergência ao fluxo. Em vez de esconder que humano e máquina chegaram a conclusões diferentes, transforma essa diferença em objeto explícito de revisão.

Isso não significa romantizar o julgamento humano. Profissionais também se ancoram, esquecem, cansam, variam entre si e podem reproduzir vieses. A razão para preservar autonomia clínica não é presumir infalibilidade humana, mas manter uma instância capaz de relacionar a recomendação a informações que o sistema pode não ter recebido, reconhecer quando o objetivo formal não corresponde ao problema do paciente e assumir responsabilidade por escolhas que afetam uma pessoa concreta. Da mesma forma, a razão para utilizar inteligência artificial não é presumir neutralidade matemática, mas explorar capacidades que humanos não possuem na mesma escala: acompanhar milhares de variáveis, comparar grandes volumes de imagens, recuperar informação, monitorar séries temporais e repetir determinadas tarefas com consistência.

Complementaridade verdadeira aparece quando os erros não se somam silenciosamente. Um algoritmo pode lembrar uma hipótese rara que o médico não considerou; o profissional pode perceber que o dado utilizado pelo sistema está desatualizado. Um modelo pode identificar deterioração em uma série temporal antes de ela se tornar visualmente evidente; a equipe pode reconhecer que o alerta se explica por uma intervenção já realizada e não exige nova ação. Um sistema generativo pode organizar centenas de páginas de prontuário; quem conhece o caso pode identificar a omissão de um evento decisivo. Nesses exemplos, o benefício não está em decidir

quem é "mais inteligente", mas em criar redundâncias úteis nas quais diferentes modos de falha se tornam mais detectáveis.

Há também uma dimensão relacional que não deve ser reduzida a uma reserva exclusiva de humanidade. Modelos conseguem produzir linguagem acolhedora, adaptar explicações e responder com aparente empatia. Isso pode ser útil, especialmente para tornar informação mais compreensível. Mas produzir uma frase empática não equivale a ocupar a posição moral e institucional de quem acompanha consequências, conhece preferências ao longo do tempo e pode ser responsabilizado por uma decisão. O vínculo clínico não é apenas uma qualidade do texto; envolve confiança construída, dever, continuidade e possibilidade de contestação. Por isso, colaboração humano-IA não deve ser desenhada para remover o paciente da relação com profissionais, mas para decidir em quais momentos a tecnologia amplia acesso e compreensão e em quais a mediação humana continua indispensável.

O cuidado aumentado, nessa perspectiva, não é aquele em que a IA participa do maior número possível de decisões. É aquele em que sua presença aumenta alguma capacidade relevante sem tornar o sistema menos seguro, menos equitativo ou menos responsável. Em alguns fluxos, isso significará automatizar uma tarefa repetitiva; em outros, oferecer segunda leitura, priorizar atenção, resumir informação ou apresentar cenários. E haverá situações em que a decisão correta será não usar o sistema. A maturidade tecnológica começa quando a ausência de IA também pode ser considerada uma escolha racional, e não um sinal de atraso.

Aprender a trabalhar com IA também é competência clínica

Se o desempenho da parceria depende da interação, capacitação deixa de ser um complemento administrativo e passa a fazer parte da segurança. Durante muito tempo, treinamento em tecnologias clínicas significou aprender a operar equipamentos ou preencher sistemas. Com modelos capazes de conversar, resumir, recomendar e produzir respostas diferentes conforme a formulação da pergunta, a competência necessária é mais ampla. Profissionais precisam compreender o que o sistema foi autorizado a fazer, que dados utiliza, em quais situações tende a falhar, como expressa incerteza, quando sua saída exige verificação e quais informações não devem ser inseridas em ferramentas não aprovadas. Isso não transforma médicos, enfermeiros, farmacêuticos ou outros profissionais em cientistas de dados. Transforma alfabetização em IA em parte da alfabetização digital exigida pelo cuidado contemporâneo.

O ensaio de Qazi et al. (2026) oferece uma evidência inicial importante porque todos os participantes realizaram vinte horas de formação sobre capacidades, limitações, estratégias de uso e avaliação crítica de modelos de linguagem antes da randomização. O estudo não permite separar perfeitamente o efeito do treinamento de todas as características do contexto, nem demonstra melhora de desfechos assistenciais; ainda assim, mostra que a preparação do usuário não pode ser presumida como constante. A mesma ferramenta pode produzir resultados diferentes quando encontra profissionais com modelos mentais diferentes sobre seu funcionamento. Esse ponto ajuda a interpretar por que estudos anteriores, realizados com pouca preparação específica, encontraram benefícios inconsistentes.

Alfabetização em IA também não deveria ser confundida com habilidade de escrever prompts sofisticados. Formular uma instrução clara é útil, mas é apenas uma pequena parte da competência. Mais importante é saber reconhecer quando o modelo não deveria ser consultado, distinguir geração de hipótese de recomendação terapêutica, verificar referências e dados factuais, perceber quando uma resposta ignora informação do prontuário e compreender que fluência verbal não corresponde a certeza. Em sistemas preditivos, é preciso interpretar probabilidades e limiares; em imagem, conhecer população e equipamento de validação; em modelos generativos, entender atualização de conhecimento, limites de contexto e risco de fabricação de informação. A competência precisa acompanhar o tipo de sistema e a consequência de sua saída.

Essa formação deve incluir o comportamento humano diante da automação. Saber que existe viés de automação não impede automaticamente que ele ocorra. Pressão de tempo, carga de trabalho, reputação do produto e modo de apresentação da interface alteram a disposição para discordar. Xu et al. (2026), ao estudarem explicações em apoio dermatológico, mostraram que a maneira como a explicação é apresentada pode modificar ancoragem e comportamento de profissionais e pessoas leigas de forma diferente. Nicolson et al. (2025) encontraram variabilidade semelhante entre clínicos expostos a explicações de um modelo. Esses resultados sugerem que treinamento não deve ensinar uma regra fixa como "confie quando houver explicação". Precisa ensinar a examinar o propósito da explicação e a evidência de que ela melhora aquela tarefa.

Também será necessário preservar competências que a automação tende a tornar menos frequentes. Quando um sistema assume parte relevante da leitura, do cálculo ou da documentação, o ganho de eficiência pode vir acompanhado de menor

exposição humana à tarefa. Isso não significa que todo profissional deva repetir manualmente aquilo que a tecnologia faz melhor. Significa que organizações precisam saber quais capacidades não podem ser perdidas porque serão necessárias diante de falha, indisponibilidade, caso atípico ou contestação. A discussão sobre deskilling não se resolve rejeitando automação, mas definindo quais competências mínimas precisam continuar praticadas, avaliadas e disponíveis.

A formação tampouco pode ficar restrita ao médico. Um sistema de IA atravessa equipes. Enfermeiros recebem alertas, farmacêuticos revisam interações, técnicos produzem imagens, profissionais administrativos lidam com agendamento e autorização, especialistas em tecnologia mantêm integrações, gestores definem indicadores e pacientes recebem mensagens ou recomendações produzidas com apoio automatizado. Se apenas um grupo entende a ferramenta, falhas podem surgir nas interfaces entre funções. A governança descrita no capítulo anterior precisa, portanto, ser traduzida em competência distribuída: cada pessoa deve conhecer aquilo que é necessário para seu papel e saber a quem recorrer quando a saída não faz sentido.

O paciente também participa dessa alfabetização, mas de maneira diferente. Não é razoável exigir que alguém compreenda arquitetura de modelos para consentir com um cuidado. É razoável, porém, oferecer informação proporcional à relevância da tecnologia: se a IA apenas auxiliou a transcrição, o conteúdo necessário será diferente daquele exigido quando um sistema selecionou quem seria priorizado para investigação ou influenciou uma proposta terapêutica. A orientação da Organização Mundial da Saúde coloca autonomia, transparência, inclusão e responsabilidade entre os princípios centrais da IA em saúde (OMS, 2021; OMS, 2024). Na prática, isso significa permitir perguntas, explicar limites importantes e preservar caminhos de revisão humana quando a decisão afeta direitos ou riscos relevantes.

Há ainda uma dimensão de aprendizagem institucional. Cada implantação produz informação sobre como profissionais realmente utilizam o sistema. Taxas de aceitação e rejeição, tipos de discordância, erros detectados pelos usuários, grupos com maior necessidade de revisão e situações em que a ferramenta é ignorada podem ensinar tanto quanto a métrica original de acurácia. Esse aprendizado não deve ser usado apenas para cobrar aderência. Discordância sistemática pode indicar falta de treinamento, mas também pode revelar que o modelo está errado, que a interface é ruim ou que a finalidade pretendida não corresponde ao fluxo real. Uma organização madura investiga a razão da divergência antes de classificá-la como resistência.

A avaliação de desempenho profissional também precisa mudar. Se um médico utiliza uma ferramenta aprovada e melhora a qualidade de uma decisão, isso não diminui sua competência; da mesma forma, rejeitar a IA por princípio não demonstra independência. A competência relevante passa a incluir capacidade de combinar fontes, reconhecer incerteza, justificar discordância e usar recursos proporcionais à dificuldade do caso. Shortliffe e Sepúlveda (2018) já defendiam que suporte à decisão deve ser pensado na interação entre sistema e usuário. A evidência atual aprofunda essa ideia: o objeto de avaliação não é apenas o que o algoritmo faz nem apenas o que o profissional sabe, mas o que o conjunto consegue produzir em condições reais.

Isso coloca uma responsabilidade nova sobre escolas, serviços de saúde e sociedades profissionais. Currículos precisam incorporar princípios de dados, avaliação crítica de modelos, segurança, equidade, privacidade e uso responsável sem transformar a formação clínica em curso de programação. Serviços precisam oferecer treinamento antes e depois da implantação, porque modelos, interfaces e políticas mudam. Normas profissionais, como a Resolução CFM nº 2.454/2026, já reconhecem capacitação e julgamento crítico como elementos do uso responsável. A formação, porém, terá valor somente se estiver ligada a sistemas concretos e a situações clínicas reais. Treinamentos genéricos sobre "o que é IA" envelhecem rápido; competências de questionar finalidade, evidência, limites e consequências permanecem.

Um futuro confiável precisa caber no sistema de saúde

A última pergunta é de escala. Uma tecnologia pode ser segura em um ensaio, útil para uma equipe treinada e ainda assim ser uma escolha ruim para um sistema de saúde. Escalar significa multiplicar licenças, integrações, processamento, armazenamento, treinamento, monitoramento, suporte e resposta a incidentes. Significa também decidir quais tarefas merecem receber recursos limitados. A sustentabilidade mencionada neste capítulo não se restringe ao ambiente, embora o inclua. Refere-se à capacidade de manter benefício clínico ao longo do tempo sem consumir mais atenção, dinheiro, infraestrutura, energia ou confiança do que o valor que a tecnologia devolve.

Essa perspectiva evita confundir inovação com acumulação. Um hospital não se torna mais avançado por possuir mais modelos. Pode, ao contrário, tornar-se mais frágil se dezenas de ferramentas competirem pela atenção dos profissionais, utilizarem integrações incompatíveis, dependerem de fornecedores sem estratégia de

continuidade e exigirem monitoramento que a instituição não consegue realizar. O princípio de proporcionalidade ao risco, presente em marcos regulatórios e no consenso FUTURE-AI, também pode ser aplicado à decisão de adoção: quanto maior a influência sobre diagnóstico, tratamento, priorização ou acesso, mais forte deve ser a evidência e maior a capacidade de supervisão (LEKADIR et al., 2025).

Sustentabilidade econômica exige observar custo total, não apenas preço de aquisição. Uma ferramenta que reduz minutos de documentação pode liberar tempo profissional, mas integração, licenciamento, validação local, segurança, atualização e auditoria também têm custo. Um algoritmo que aumenta detecção pode gerar exames adicionais, encaminhamentos e tratamentos que precisam ser absorvidos pela rede. Em outro contexto, uma solução relativamente simples pode ter alto valor se ampliar rastreamento onde não havia especialista disponível. Como discutimos desde o Capítulo 2, o comparador correto não é sempre o melhor especialista em condições ideais; é a alternativa real existente naquele sistema.

A mesma lógica vale para sustentabilidade ambiental. A medicina digital depende de centros de dados, redes, armazenamento, equipamentos e cadeias de hardware que consomem energia, água e materiais. Schmitz et al. (2026) chamam atenção para a necessidade de equilibrar benefícios clínicos da IA com carbono, uso de recursos naturais e resíduos eletrônicos. A questão é particularmente relevante para modelos de grande escala, nos quais a escolha entre arquiteturas, frequência de uso e infraestrutura pode alterar substancialmente a demanda computacional. Isso não significa atribuir a cada consulta um cálculo impossível de pegada ambiental, mas incorporar eficiência computacional e impacto ambiental entre os critérios de desenvolvimento e compra, sobretudo quando duas soluções oferecem benefício clínico semelhante.

O problema ainda é pouco medido. Nooruddin et al. (2026), em revisão de escopo sobre modelos generativos em saúde, identificaram 24 estudos elegíveis e observaram que apenas um relatava a pegada de carbono. A lacuna é instrutiva: o campo desenvolveu vocabulário sofisticado para acurácia, validação e segurança, mas ainda mede de forma incipiente o custo ambiental de produzir essas capacidades. Uma saúde que se pretende sustentável não pode tratar o impacto computacional como externalidade invisível, especialmente quando o próprio setor saúde já participa de emissões e consumo relevantes de recursos.

Sustentabilidade também é social. Uma tecnologia que só funciona com conectividade de alta qualidade, equipamentos recentes, documentação padronizada em determinado idioma ou equipes capazes de pagar treinamento especializado pode ampliar distância entre serviços. O Capítulo 7 mostrou que desigualdade pode estar dentro dos dados; ela também pode estar na infraestrutura necessária para usar o modelo. A revisão de Wu et al. (2026) sobre exclusão digital em pessoas idosas lembra que acesso a tecnologias de saúde é mediado por letramento, recursos, suporte e condições sociais. Portanto, ampliar IA sem planejar acesso pode criar uma forma nova de diferença: alguns pacientes recebem cuidado aumentado por informação digital abundante, enquanto outros permanecem pouco representados e pouco alcançados.

É por isso que sistemas de saúde precisam aprender também a desimplantar. A adoção costuma receber mais atenção do que a retirada, mas tecnologias que não entregam benefício, acumulam incidentes, perderam validade ou foram superadas por alternativas mais simples não deveriam permanecer apenas porque já foram integradas. Critérios de suspensão discutidos no Capítulo 9 são parte da sustentabilidade: permitem interromper uma solução sem transformar cada retirada em fracasso político ou financeiro. Inovação responsável inclui a capacidade de encerrar o que não funciona.

Esse princípio se aplica igualmente ao tamanho dos modelos. Nem toda tarefa clínica precisa da arquitetura mais geral ou mais poderosa disponível. Classificadores especializados, regras bem construídas, calculadoras validadas ou modelos menores podem ser mais transparentes, baratos e suficientes para certas finalidades. Modelos fundacionais têm valor justamente pela flexibilidade, mas flexibilidade sem necessidade pode aumentar complexidade e superfície de risco. A pergunta de engenharia deve permanecer subordinada à pergunta clínica: qual é a menor complexidade capaz de produzir o benefício necessário com segurança aceitável? Esse critério aproxima eficiência técnica de prudência assistencial.

CONSIDERAÇÕES FINAIS

Ao longo deste livro, a inteligência artificial apareceu como classificador, leitor de imagem, preditor, sistema de alerta, ferramenta de personalização, sensor, escriba, modelo generativo e componente de governança. Nenhuma dessas formas possui valor isoladamente. O que importa é a mudança produzida no percurso de cuidado. Um diagnóstico mais precoce só é benefício se houver caminho para confirmação e tratamento; um alerta só ajuda se alguém puder responder; uma recomendação personalizada só é legítima se sua evidência for causalmente defensável; uma resposta generativa só economiza tempo se a verificação não consumir mais do que foi poupado; uma ferramenta equitativa precisa continuar equitativa depois da implantação.

Essa sequência ajuda a definir o que "cuidado aumentado" realmente significa. Não é um cuidado em que a máquina ocupa mais espaço. É um cuidado em que capacidades técnicas ampliam percepção, organização e antecipação sem reduzir autonomia, vínculo e responsabilidade. O aumento pode aparecer como acesso a rastreamento, recuperação de informação esquecida, redução de trabalho documental, identificação de deterioração ou possibilidade de comparar cenários que seriam difíceis de integrar manualmente. Mas só merece esse nome quando o ganho permanece visível depois que consideramos falsos positivos, vieses, carga de trabalho, privacidade, segurança, custos e efeitos sobre a relação clínica.

O futuro mais confiável, portanto, talvez seja menos espetacular do que algumas previsões sugerem. Em vez de um sistema universal substituindo equipes, é provável que a inteligência artificial se distribua por centenas de pequenas decisões: algumas invisíveis, outras explicitamente compartilhadas com profissionais e pacientes. Haverá ferramentas autônomas em tarefas delimitadas, sistemas de segunda leitura, agentes capazes de executar partes de fluxos e modelos que funcionam como interlocutores. A diversidade de funções torna ainda menos plausível uma regra única para "uso de IA". Cada aplicação precisará justificar seu lugar pela relação entre benefício, risco, alternativas e capacidade de governança.

Essa conclusão também muda a pergunta sobre confiança. Não precisamos de profissionais ou pacientes que confiem mais em inteligência artificial. Precisamos de sistemas que sejam dignos da confiança exigida para a tarefa e de pessoas capazes de ajustar sua dependência quando a evidência muda. Confiança não é entusiasmo. É

uma expectativa sustentada por validação, transparência suficiente, possibilidade de contestação, monitoramento e responsabilidade. Quando esses elementos faltam, pedir confiança é substituir evidência por persuasão.

No início da obra, a questão era o que significa permitir que uma máquina participe de atividades historicamente associadas ao julgamento clínico. Depois de percorrer diagnóstico, previsão, monitoramento, geração de linguagem, validação, vieses, privacidade, segurança e responsabilidade, a resposta é mais precisa. Participar não é assumir o cuidado. É ocupar uma função delimitada dentro de uma rede de decisões humanas e institucionais. Quanto maior essa função, maior deve ser a evidência que a sustenta e mais nítidos precisam ser os caminhos de supervisão e reparação.

A inteligência artificial mais valiosa para a saúde não será necessariamente a que parecer mais humana, produzir a resposta mais fluente ou alcançar a maior pontuação em um benchmark. Será a que conseguir permanecer útil quando encontrar pacientes diferentes, profissionais cansados, dados incompletos, serviços desiguais e consequências que não cabem em uma métrica única. Será também aquela que puder ser questionada, monitorada, corrigida e, quando necessário, retirada. A excelência não está em eliminar incerteza da medicina, mas em evitar que a tecnologia a esconda.

Por isso, o ponto de chegada deste livro não é uma defesa da automação nem uma defesa do julgamento humano isolado. É uma defesa de sistemas de cuidado capazes de usar inteligência artificial sem terceirizar aquilo que torna o cuidado responsável: reconhecer limites, responder por consequências, proteger quem está vulnerável e revisar decisões quando a realidade contradiz o modelo. Se a tecnologia ampliar essas capacidades, teremos uma medicina aumentada no sentido mais importante da expressão. Se as reduzir, mesmo um algoritmo tecnicamente extraordinário terá produzido pouco progresso clínico.

REFERÊNCIAS

ABRÀMOFF, Michael D.; LAVIN, Philip T.; BIRCH, Michele; SHAH, Nilay; FOLK, James C. Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. npj Digital Medicine, v. 1, art. 39, 2018. DOI: 10.1038/s41746-018-0040-6.

ACOSTA, Julián N.; FALCONE, Guido J.; RAJPURKAR, Pranav; TOPOL, Eric J. Multimodal biomedical AI. Nature Medicine, v. 28, p. 1773-1784, 2022. DOI: 10.1038/s41591-022-01981-2.

ADAMS, Roy J. et al. Prospective, multi-site study of patient outcomes after implementation of the TREWS machine learning-based early warning system for sepsis. Nature Medicine, v. 28, p. 1455-1460, 2022. DOI: 10.1038/s41591-022-01894-0.

AGÊNCIA NACIONAL DE VIGILÂNCIA SANITÁRIA (ANVISA). Agenda Regulatória 2026-2027: lista final de temas. Brasília, DF: Anvisa, 2026. Disponível em: https://www.gov.br/anvisa/pt-br/assuntos/regulamentacao/agenda-regulatoria/agenda-2026-2027. Acesso em: 21 ago. 2026.

AGÊNCIA NACIONAL DE VIGILÂNCIA SANITÁRIA (ANVISA). Resolução da Diretoria Colegiada - RDC nº 657, de 24 de março de 2022. Dispõe sobre a regularização de software como dispositivo médico (Software as a Medical Device - SaMD). Brasília, DF: Anvisa, 2022a.

AGÊNCIA NACIONAL DE VIGILÂNCIA SANITÁRIA (ANVISA). Software como dispositivo médico: perguntas e respostas sobre a RDC nº 657/2022. Brasília, DF: Anvisa, 2022b. Disponível em: https://www.gov.br/anvisa/pt-br/centraisdeconteudo/publicacoes/produtos-para-a-saude/manuais/software-como-dispositivo-medico-perguntas-e-respostas. Acesso em: 21 ago. 2026.

AHN, J. S. et al. Association of artificial intelligence-aided chest radiograph interpretation with reader performance and efficiency. JAMA Network Open, v. 5, n. 8, e2229289, 2022. DOI: 10.1001/jamanetworkopen.2022.29289.

AMANN, Julia; BLASIMME, Alessandro; VAYENA, Effy; FREY, Dietmar; MADAI, Vince I. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC Medical Informatics and Decision Making, v. 20, art. 310, 2020. DOI: 10.1186/s12911-020-01332-6.

AUTORIDADE NACIONAL DE PROTEÇÃO DE DADOS (ANPD). Resolução CD/ANPD nº 15, de 24 de abril de 2024. Aprova o Regulamento de Comunicação de Incidente de Segurança. Brasília, DF: ANPD, 2024.

AZAD, Tej D.; KRUMHOLZ, Harlan M.; SARIA, Suchi. Principles to guide clinical AI readiness and move from benchmarks to real-world evaluation. Nature Medicine, v. 32, p. 802-804, 2026. DOI: 10.1038/s41591-025-04198-1.

BOTERO, Dany Alexandra Morales et al. Artificial intelligence in Brazilian Primary Health Care: scoping review. Revista Brasileira de Enfermagem, v. 78, supl. 3, e20240363, 2025. DOI: 10.1590/0034-7167-2024-0363.

BOUHOUITA-GUERMECH, Sarah; HAIDAR, Hazar. Scoping review shows the dynamics and complexities inherent to the notion of "responsibility" in artificial intelligence within the healthcare context. Asian Bioethics Review, v. 16, n. 3, p. 315-344, 2024. DOI: 10.1007/s41649-024-00292-7.

BRASIL. Câmara dos Deputados. Projeto de Lei nº 2.338, de 2023. Dispõe sobre o desenvolvimento, o fomento e o uso ético e responsável da inteligência artificial com base na centralidade da pessoa humana. Brasília, DF: Câmara dos Deputados, 2026. Disponível em: https://www.camara.leg.br/proposicoesWeb/fichadetramitacao?idProposicao=2487262. Acesso em: 22 ago. 2026.

BRASIL. Decreto nº 12.560, de 23 de julho de 2025. Dispõe sobre a Rede Nacional de Dados em Saúde e sobre as Plataformas SUS Digital e regulamenta o art. 47 e o art. 47-A da Lei nº 8.080, de 19 de setembro de 1990. Diário Oficial da União, Brasília, DF, 24 jul. 2025.

BRASIL. Lei nº 13.709, de 14 de agosto de 2018. Lei Geral de Proteção de Dados Pessoais (LGPD). Diário Oficial da União, Brasília, DF, 15 ago. 2018.

BUSCH, Felix et al. Current applications and challenges in large language models for patient care: a systematic review. Communications Medicine, v. 5, art. 26, 2025. DOI: 10.1038/s43856-024-00717-2.

CESTONARO, Clara; DELICATI, Arianna; MARCANTE, Beatrice; CAENAZZO, Luciana; TOZZO, Pamela. Defining medical liability when artificial intelligence is applied on diagnostic algorithms: a systematic review. Frontiers in Medicine, v. 10, art. 1305756, 2023. DOI: 10.3389/fmed.2023.1305756.

COLLINS, Gary S. et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ, v. 385, e078378, 2024. DOI: 10.1136/bmj-2023-078378.

COMEAU, Donnella S.; BITTERMAN, Danielle S.; CELI, Leo Anthony. Preventing unrestricted and unmonitored AI experimentation in healthcare through transparency and accountability. npj Digital Medicine, v. 8, art. 42, 2025. DOI: 10.1038/s41746-025-01443-2.

CONSELHO FEDERAL DE MEDICINA (CFM). Resolução CFM nº 2.454, de 11 de fevereiro de 2026. Normatiza o uso da inteligência artificial na medicina. Publicada no Diário Oficial da União em 27 fev. 2026; retificação em 5 mar. 2026. Brasília, DF: CFM, 2026. Disponível em: https://sistemas.cfm.org.br/normas/visualizar/resolucoes/br/2026/2454. Acesso em: 22 ago. 2026.

CRUZ RIVERA, Samantha et al. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI Extension. BMJ, v. 370, m3210, 2020. DOI: 10.1136/bmj.m3210.

EISEMANN, Nora et al. Nationwide real-world implementation of AI for cancer detection in population-based mammography screening. Nature Medicine, v. 31, p. 917-924, 2025. DOI: 10.1038/s41591-024-03408-6.

ESTEVA, Andre et al. A guide to deep learning in healthcare. Nature Medicine, v. 25, p. 24-29, 2019. DOI: 10.1038/s41591-018-0316-z.

EVERETT, Selin S. et al. From tool to teammate in a randomized controlled trial of clinician-AI collaborative workflows for diagnosis. npj Digital Medicine, v. 9, art. 409, 2026. DOI: 10.1038/s41746-026-02545-1.

FOOD AND DRUG ADMINISTRATION (FDA). Artificial Intelligence-Enabled Medical Devices. Silver Spring, MD: U.S. Food and Drug Administration, 2026a. Disponível em: https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-enabled-medical-devices. Acesso em: 21 ago. 2026.

FOOD AND DRUG ADMINISTRATION (FDA). Clinical Decision Support Software: Guidance for Industry and Food and Drug Administration Staff. Silver Spring, MD: U.S. Food and Drug Administration, 2026b. Disponível em: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/clinical-decision-support-software. Acesso em: 22 ago. 2026.

FOOD AND DRUG ADMINISTRATION (FDA). Cybersecurity in Medical Devices: Quality Management System Considerations and Content of Premarket Submissions. Guidance for Industry and Food and Drug Administration Staff. Silver Spring, MD: FDA, 2026c. Disponível em: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/cybersecurity-medical-devices-quality-management-system-considerations-and-content-premarket. Acesso em: 22 ago. 2026.

FOOD AND DRUG ADMINISTRATION (FDA). Marketing Submission Recommendations for a Predetermined Change Control Plan for Artificial Intelligence-Enabled Device Software Functions: Guidance for Industry and Food and Drug Administration Staff. Silver Spring, MD: FDA, 2025. Disponível em: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/marketing-submission-recommendations-predetermined-change-control-plan-artificial-intelligence. Acesso em: 22 ago. 2026.

FOOD AND DRUG ADMINISTRATION (FDA); HEALTH CANADA; MEDICINES AND HEALTHCARE PRODUCTS REGULATORY AGENCY (MHRA). Transparency for Machine Learning-Enabled Medical Devices: Guiding Principles. 2024. Disponível em: https://www.fda.gov/medical-devices/software-medical-device-samd/transparency-machine-learning-enabled-medical-devices-guiding-principles. Acesso em: 22 ago. 2026.

FOTHERINGHAM, Kit; SMITH, Helen. Accidental injustice: healthcare AI legal responsibility must be prospectively planned prior to its adoption. Future Healthcare Journal, v. 11, n. 3, art. 100181, 2024. DOI: 10.1016/j.fhj.2024.100181.

GALLIFANT, Jack et al. The TRIPOD-LLM reporting guideline for studies using large language models. Nature Medicine, v. 31, p. 60-69, 2025. DOI: 10.1038/s41591-024-03425-5.

GHASSEMI, Marzyeh; OAKDEN-RAYNER, Luke; BEAM, Andrew L. The false hope of current approaches to explainable artificial intelligence in health care. The Lancet Digital Health, v. 3, n. 11, p. e745-e750, 2021. DOI: 10.1016/S2589-7500(21)00208-9.

GICHOYA, Judy Wawira et al. AI recognition of patient race in medical imaging: a modelling study. The Lancet Digital Health, v. 4, n. 6, p. e406-e414, 2022. DOI: 10.1016/S2589-7500(22)00063-2.

GODDARD, Kate; ROUDSARI, Abdul; WYATT, Jeremy C. Automation bias: a systematic review of frequency, effect mediators, and mitigators. Journal of the American Medical Informatics Association, v. 19, n. 1, p. 121-127, 2012. DOI: 10.1136/amiajnl-2011-000089.

GOH, Ethan et al. GPT-4 assistance for improvement of physician performance on patient care tasks: a randomized controlled trial. Nature Medicine, v. 31, p. 1233-1238, 2025. DOI: 10.1038/s41591-024-03456-y.

GOH, Ethan et al. Large language model influence on diagnostic reasoning: a randomized clinical trial. JAMA Network Open, v. 7, n. 10, e2440969, 2024. DOI: 10.1001/jamanetworkopen.2024.40969.

GOMMERS, Jessie et al. Interval cancer, sensitivity, and specificity comparing AI-supported mammography screening with standard double reading without AI in the MASAI study: a randomised, controlled, non-inferiority, single-blinded, population-based, screening-accuracy trial. The Lancet, v. 407, n. 10527, p. 505-514, 2026. DOI: 10.1016/S0140-6736(25)02464-X.

GROH, Matthew et al. Deep learning-aided decision support for diagnosis of skin disease across skin tones. Nature Medicine, v. 30, p. 573-583, 2024. DOI: 10.1038/s41591-023-02728-3.

HAGER, Paul et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. Nature Medicine, v. 30, p. 2613-2622, 2024. DOI: 10.1038/s41591-024-03097-1.

HENRY, Katharine E. et al. Factors driving provider adoption of the TREWS machine learning-based early warning system and its effects on sepsis treatment timing. Nature Medicine, v. 28, p. 1447-1454, 2022. DOI: 10.1038/s41591-022-01895-z.

HERNSTRÖM, Veronica et al. Screening performance and characteristics of breast cancer detected in the Mammography Screening with Artificial Intelligence trial (MASAI): a randomised, controlled, parallel-group, non-inferiority, single-blinded, screening accuracy study. The Lancet Digital Health, v. 7, n. 3, p. e175-e183, 2025. DOI: 10.1016/S2589-7500(24)00267-X.

INTERNATIONAL MEDICAL DEVICE REGULATORS FORUM (IMDRF). Good machine learning practice for medical device development: guiding principles. IMDRF/AIML WG/N88 FINAL:2025. 29 jan. 2025. Disponível em: https://www.imdrf.org/documents/good-machine-learning-practice-medical-device-development-guiding-principles. Acesso em: 22 ago. 2026.

KAISSIS, Georgios A.; MAKOWSKI, Marcus R.; RÜCKERT, Daniel; BRAREN, Rickmer F. Secure, privacy-preserving and federated machine learning in medical imaging. Nature Machine Intelligence, v. 2, p. 305-311, 2020. DOI: 10.1038/s42256-020-0186-1.

KAUL, Vivek; ENSLIN, Sarah; GROSS, Seth A. History of artificial intelligence in medicine. Gastrointestinal Endoscopy, v. 92, n. 4, p. 807-812, 2020. DOI: 10.1016/j.gie.2020.06.040.

KENT, David M.; STEYERBERG, Ewout; VAN KLAVEREN, David. Personalized evidence based medicine: predictive approaches to heterogeneous treatment effects. BMJ, v. 363, k4245, 2018. DOI: 10.1136/bmj.k4245.

KOLLA, Likhitha et al. Performance drift in a nationally deployed population health risk algorithm in the US Veterans Health Administration. JAMA Health Forum, v. 6, n. 8, e252717, 2025. DOI: 10.1001/jamahealthforum.2025.2717.

KOMOROWSKI, Matthieu et al. The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care. Nature Medicine, v. 24, p. 1716-1720, 2018. DOI: 10.1038/s41591-018-0213-5.

KORE, Ali et al. Empirical data drift detection experiments on real-world medical imaging data. Nature Communications, v. 15, art. 1887, 2024. DOI: 10.1038/s41467-024-46142-w.

LEKADIR, Karim et al. FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare. BMJ, v. 388, e081554, 2025. DOI: 10.1136/bmj-2024-081554.

LIU, Xiaohong et al. A generalist medical language model for disease diagnosis assistance. Nature Medicine, v. 31, p. 932-942, 2025. DOI: 10.1038/s41591-024-03416-6.

LIU, Xiaoxuan et al. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI Extension. BMJ, v. 370, m3164, 2020. DOI: 10.1136/bmj.m3164.

LU, Ming Y. et al. A multimodal generative AI copilot for human pathology. Nature, v. 634, p. 466-473, 2024. DOI: 10.1038/s41586-024-07618-3.

LUKAC, Paul J. et al. Ambient AI scribes in clinical practice: a randomized trial. NEJM AI, v. 2, n. 12, 2025. DOI: 10.1056/aioa2501000.

MILLER, Randolph A.; POPLE, Harry E.; MYERS, Jack D. INTERNIST-1, an experimental computer-based diagnostic consultant for general internal medicine. New England Journal of Medicine, v. 307, n. 8, p. 468-476, 1982. DOI: 10.1056/NEJM198208193070803.

MOONS, Karel G. M. et al. PROBAST+AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. BMJ, v. 388, e082505, 2025. DOI: 10.1136/bmj-2024-082505.

MOOR, Michael et al. Foundation models for generalist medical artificial intelligence. Nature, v. 616, p. 259-265, 2023. DOI: 10.1038/s41586-023-05881-4.

NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY (NIST). Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations. NIST AI 100-2e2025. Gaithersburg, MD: NIST, 2025. DOI: 10.6028/NIST.AI.100-2e2025.

NICOLSON, Angus et al. The human factor in explainable artificial intelligence: clinician variability in trust, reliance, and performance. npj Digital Medicine, v. 8, art. 658, 2025. DOI: 10.1038/s41746-025-02023-0.

NOORUDDIN, Ayesha et al. Use of generative artificial intelligence models in healthcare - a scoping review. BMC Artificial Intelligence, v. 2, art. 11, 2026. DOI: 10.1186/s44398-026-00029-6.

OBERMEYER, Ziad; EMANUEL, Ezekiel J. Predicting the future - big data, machine learning, and clinical medicine. New England Journal of Medicine, v. 375, n. 13, p. 1216-1219, 2016. DOI: 10.1056/NEJMp1606181.

OBERMEYER, Ziad; POWERS, Brian; VOGELI, Christine; MULLAINATHAN, Sendhil. Dissecting racial bias in an algorithm used to manage the health of populations. Science, v. 366, n. 6464, p. 447-453, 2019. DOI: 10.1126/science.aax2342.

OMAR, Mahmud et al. Sociodemographic biases in medical decision making by large language models. Nature Medicine, v. 31, p. 1873-1881, 2025. DOI: 10.1038/s41591-025-03626-6.

ORGANIZAÇÃO MUNDIAL DA SAÚDE (OMS). Ethics and governance of artificial intelligence for health: guidance on large multi-modal models. Geneva: World Health Organization, 2024. ISBN: 978-92-4-008475-9.

ORGANIZAÇÃO MUNDIAL DA SAÚDE (OMS). Ethics and governance of artificial intelligence for health: WHO guidance. Geneva: World Health Organization, 2021. ISBN: 978-92-4-002920-0.

PEREZ, Marco V. et al. Large-scale assessment of a smartwatch to identify atrial fibrillation. New England Journal of Medicine, v. 381, n. 20, p. 1909-1917, 2019. DOI: 10.1056/NEJMoa1901183.

PFOHL, Stephen R. et al. A toolbox for surfacing health equity harms and biases in large language models. Nature Medicine, v. 30, p. 3590-3600, 2024. DOI: 10.1038/s41591-024-03258-2.

PRICE II, W. Nicholson; COHEN, I. Glenn. Privacy in the age of medical big data. Nature Medicine, v. 25, p. 37-43, 2019. DOI: 10.1038/s41591-018-0272-7.

QAZI, Ihsan Ayyub et al. Large language model diagnostic assistance for physicians in a lower-middle-income country: a randomized controlled trial. Nature Health, v. 1, p. 198-205, 2026. DOI: 10.1038/s44360-025-00007-8.

RAJKOMAR, Alvin et al. Scalable and accurate deep learning with electronic health records. npj Digital Medicine, v. 1, art. 18, 2018. DOI: 10.1038/s41746-018-0029-1.

RAJPURKAR, Pranav et al. AI in health and medicine. Nature Medicine, v. 28, p. 31-38, 2022. DOI: 10.1038/s41591-021-01614-0.

RALDOW, Ann C. et al. Proactive care management of AI-identified at-risk patients decreases preventable admissions. American Journal of Managed Care, v. 30, n. 11, p. 548-554, 2024. DOI: 10.37765/ajmc.2024.89625.

RIEKE, Nicola et al. The future of digital health with federated learning. npj Digital Medicine, v. 3, art. 119, 2020. DOI: 10.1038/s41746-020-00323-1.

SCHMITZ, Nikolas E. J. et al. Sustainable AI in medicine: balancing benefits and environmental costs. Nature Reviews Nephrology, v. 22, p. 165-166, 2026. DOI: 10.1038/s41581-026-01047-3.

SEYYED-KALANTARI, Laleh et al. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. Nature Medicine, v. 27, p. 2176-2182, 2021. DOI: 10.1038/s41591-021-01595-0.

SHOKRI, Reza; STRONATI, Marco; SONG, Congzheng; SHMATIKOV, Vitaly. Membership inference attacks against machine learning models. In: IEEE SYMPOSIUM ON SECURITY AND PRIVACY, 2017. Proceedings [...]. IEEE, 2017. p. 3-18. DOI: 10.1109/SP.2017.41.

SHORTLIFFE, Edward H.; SEPULVEDA, Martin J. Clinical decision support in the era of artificial intelligence. JAMA, v. 320, n. 21, p. 2199-2200, 2018. DOI: 10.1001/jama.2018.17163.

SIDERIS, Konstantinos et al. Artificial intelligence predictive analytics in heart failure: results of the pilot phase of a pragmatic randomized clinical trial. Journal of the American Medical Informatics Association, v. 31, n. 4, p. 919-928, 2024. DOI: 10.1093/jamia/ocae017.

SIDERIS, Konstantinos et al. Intervention based on remote monitoring and artificial intelligence enabled predictive analytics in heart failure: the LINK-HF2 multicenter randomized trial. Journal of the American College of Cardiology, v. 87, n. 13 Supplement, p. A554, 2026. DOI: 10.1016/j.jacc.2026.02.1489.

SINGHAL, Karan et al. Large language models encode clinical knowledge. Nature, v. 620, p. 172-180, 2023. DOI: 10.1038/s41586-023-06291-2.

SOUNDERAJAH, Viknesh et al. The STARD-AI reporting guideline for diagnostic accuracy studies using artificial intelligence. Nature Medicine, v. 31, p. 3283-3289, 2025. DOI: 10.1038/s41591-025-03953-8.

TAI-SEALE, Ming et al. AI-generated draft replies integrated into health records and physicians' electronic communication. JAMA Network Open, v. 7, n. 4, e246565, 2024. DOI: 10.1001/jamanetworkopen.2024.6565.

TIKHOMIROV, Lana et al. A scoping review of silent trials for medical artificial intelligence. Nature Health, v. 1, p. 532-554, 2026. DOI: 10.1038/s44360-025-00048-z.

TOMAŠEV, Nenad et al. A clinically applicable approach to continuous prediction of future acute kidney injury. Nature, v. 572, n. 7767, p. 116-119, 2019. DOI: 10.1038/s41586-019-1390-1.

TOPOL, Eric J. High-performance medicine: the convergence of human and artificial intelligence. Nature Medicine, v. 25, p. 44-56, 2019. DOI: 10.1038/s41591-018-0300-7.

TURING, Alan M. Computing machinery and intelligence. Mind, v. 59, n. 236, p. 433-460, 1950. DOI: 10.1093/mind/LIX.236.433.

UNIÃO EUROPEIA. Regulamento (UE) 2024/1689 do Parlamento Europeu e do Conselho, de 13 de junho de 2024, que cria regras harmonizadas em matéria de inteligência artificial (Regulamento da Inteligência Artificial). Jornal Oficial da União Europeia, 12 jul. 2024. Texto consolidado em 27 jul. 2026. Disponível em: https://eur-lex.europa.eu/eli/reg/2024/1689/2026-07-27/eng. Acesso em: 22 ago. 2026.

VAN AMSTERDAM, Wouter A. C. et al. Clinical trials for continuously monitored and updated AI systems. Nature Medicine, v. 32, p. 1972-1974, 2026. DOI: 10.1038/s41591-026-04368-9.

VAN STEIJN, Nicole J. et al. Enhanced detection and prompt diagnosis of atrial fibrillation using Apple Watch: a randomized controlled trial. Journal of the American College of Cardiology, v. 87, n. 14, p. 1714-1728, 2026. DOI: 10.1016/j.jacc.2025.11.032.

VAN VEEN, Dave et al. Adapted large language models can outperform medical experts in clinical text summarization. Nature Medicine, v. 30, p. 1134-1142, 2024. DOI: 10.1038/s41591-024-02855-5.

VASEY, Baptiste et al. Reporting guideline for the early stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. BMJ, v. 377, e070904, 2022. DOI: 10.1136/bmj-2022-070904.

WANG, Ting-Wei et al. Systematic review and meta-analysis of regulator-approved deep learning systems for fundus diabetic retinopathy detections. npj Digital Medicine, v. 9, art. 110, 2026. DOI: 10.1038/s41746-025-02223-8.

WOLBER, Jonas et al. Predicting disease outcomes from remote monitoring using machine learning: a systematic review. BMC Medical Informatics and Decision Making, v. 26, art. 145, 2026. DOI: 10.1186/s12911-026-03495-0.

WONG, Andrew et al. External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. JAMA Internal Medicine, v. 181, n. 8, p. 1065-1070, 2021. DOI: 10.1001/jamainternmed.2021.2626.

WORLD HEALTH ORGANIZATION (WHO). Addressing future cybersecurity threats in digital health: report of the technical consultation, Geneva, Switzerland, 12-14 December 2023. Geneva: World Health Organization, 2024. ISBN: 978-92-4-009926-5.

WU, Yucen et al. Digital exclusion and health in older adults: a systematic review. npj Digital Medicine, 2026. DOI: 10.1038/s41746-026-03091-6.

XU, Xuhai Orson et al. Divergent impacts of explainable AI for dermatological diagnosis on clinicians versus lay people. Nature Medicine, v. 32, p. 3000-3009, 2026. DOI: 10.1038/s41591-026-04553-w.

YANG, Yuzhe et al. The limits of fair medical imaging AI in real-world generalization. Nature Medicine, v. 30, p. 2838-2848, 2024. DOI: 10.1038/s41591-024-03113-4.

YENDE, Sachin et al. Remote monitoring approaches to reduce readmissions after infection and sepsis: a randomized clinical trial. JAMA Network Open, v. 9, n. 6, e2616641, 2026. DOI: 10.1001/jamanetworkopen.2026.16641.

YU, Kun-Hsing; BEAM, Andrew L.; KOHANE, Isaac S. Artificial intelligence in healthcare. Nature Biomedical Engineering, v. 2, p. 719-731, 2018. DOI: 10.1038/s41551-018-0305-z.

ZHANG, Kai et al. A generalist vision-language foundation model for diverse biomedical tasks. Nature Medicine, v. 30, p. 3129-3141, 2024. DOI: 10.1038/s41591-024-03185-2.

ZHANG, S. et al. An AI-based OCT system to detect diabetic macular edema: a prospective validation and noninferiority randomized clinical trial. JAMA, v. 336, n. 3, p. 215-223, 2026. DOI: 10.1001/jama.2026.7025.

SOBRE OS AUTORES

Flávio Eduardo Fernandes Alves

Flávio Eduardo Fernandes Alves é graduado em Biomedicina e Ciências Biológicas, especialista em Gestão Pública em Saúde e mestrando em Ciência e Tecnologia Ambiental. É servidor público federal da Universidade Federal do Triângulo Mineiro, atuando como Técnico em Anatomia e Necropsia no Hospital de Clínicas. Sua experiência profissional e acadêmica envolve Anatomia, Necropsia, Patologia Cirúrgica, Saúde, Gestão Pública e Ciência e Tecnologia Ambiental, favorecendo abordagens interdisciplinares sobre cuidado, gestão e sustentabilidade.

Ana Carla do Espírito Santo

Ana Carla do Espírito Santo é enfermeira graduada pela Universidade Federal do Estado do Rio de Janeiro (UNIRIO) e possui formação de pós-graduação em áreas da enfermagem. Atua como enfermeira estatutária da Secretaria de Estado de Saúde do Distrito Federal, com experiência na Atenção Primária à Saúde e na Estratégia Saúde da Família. Desenvolve interesse em saúde da mulher, planejamento reprodutivo, enfermagem de família e comunidade, educação permanente em saúde e qualificação do cuidado.

Márcia Silva Conceição

Márcia Silva Conceição é enfermeira graduada pelas Faculdades Integradas do Tapajós, mestra em Ciência da Saúde pela Universidade Federal de Roraima e pós-graduada em Enfermagem Obstétrica pela Universidade Estadual de Roraima. É servidora estatutária municipal em Boa Vista e atua como enfermeira na Atenção Básica. Possui experiência em Saúde da Família, Saúde da Mulher e Saúde Indígena, com trajetória profissional voltada ao cuidado materno, à atenção territorial e à assistência integral em saúde.

José Peixoto de Alencar Neto

José Peixoto de Alencar Neto é médico formado pela Universidade Tiradentes, com pós-graduação em Psiquiatria pela Universidade São Judas Tadeu e em Saúde Mental, além de formação em andamento em Nutrologia pela Universidade de São Paulo – Ribeirão Preto. Sua produção acadêmica abrange temas de saúde mental,

transtorno de estresse pós-traumático, condições cardiovasculares e manifestações neurológicas, articulando prática médica, atualização profissional e investigação científica.



INTELIGÊNCIA ARTIFICIAL NA SAÚDE

A inteligência artificial deixou de ser promessa e passou a integrar o cuidado: reconhece padrões, estima riscos, interpreta imagens, organiza informação clínica e produz linguagem. A pergunta que esta obra enfrenta não é se a IA pode participar da assistência, mas sob quais condições essa participação produz benefício clínico real sem ampliar riscos, desigualdades e zonas de responsabilidade indefinida.

Em dez capítulos, o livro percorre o caminho da lógica simbólica aos modelos generativos e multimodais e examina diagnóstico por imagem, prognóstico, medicina personalizada, monitoramento remoto e saúde populacional. Dedica atenção especial à distância entre acurácia experimental e segurança clínica: validação externa, viés algorítmico, desempenho desigual entre grupos e risco de automação excessiva.

Privacidade, explicabilidade, cibersegurança, regulação e responsabilidade profissional não aparecem como etapa burocrática posterior, e sim como parte do próprio desenho do cuidado. A perspectiva sustentada é a do cuidado aumentado, e não substituído: tecnologias que ampliam capacidades sem dissolver julgamento profissional, vínculo, equidade e prestação de contas.

Obra destinada a profissionais e estudantes das áreas da saúde, pesquisadores, gestores e equipes envolvidas com transformação digital.

Organização: Kátia Karolini Amaro El Alam

EDITORA MANUAL

www.editoramanual.com.br
1ª edição • 2026

ISBN: 978-65-83121-17-2