noemicajournal.org

e-ISSN: 3086-7037

ARTIGO

Governança responsável da inteligência artificial em sistemas de saúde: princípios para equidade, transparência e capacidade institucional

Responsible governance of artificial intelligence in healthcare systems: principles for equity, transparency, and institutional capacity

Gobernanza responsable de la inteligencia artificial en sistemas de salud: principios para la equidad, la transparencia y la capacidad institucional

Antonio Pires Barbosa[1]

[1] Universidade Nove de Julho (UNINOVE), São Paulo, SP, Brasil. ORCID: https://orcid.org/0000-0001-6478-6522.

SUBMETIDO	ACEITO	PUBLICADO	DOI
12/03/2025	20/10/2025	05/12/2025	em processamento

RESUMO

A incorporação da inteligência artificial em sistemas de saúde levanta desafios de governança que ultrapassam a validação técnica dos algoritmos. Este artigo examina como princípios normativos podem orientar a adoção responsável de sistemas de diagnóstico, triagem e gestão clínica, com atenção especial à equidade no atendimento, à transparência das decisões automatizadas e ao fortalecimento da capacidade institucional das organizações de saúde. A discussão parte de uma revisão integrativa da literatura sobre governança clínica, segurança do paciente, justiça distributiva, explicabilidade algorítmica e capacidades organizacionais, articulando contribuições de periódicos como Nature Medicine, The New England Journal of Medicine, Science, BMC Medicine e das diretrizes da Organização Mundial da Saúde sobre ética e governança da inteligência artificial para a saúde. A partir dessa base, propõe-se um modelo com seis dimensões inter-relacionadas: finalidade clínica legítima, qualidade e representatividade dos dados, validação contextual, supervisão humana efetiva, mecanismos de contestação e monitoramento contínuo. Argumenta-

se que o desempenho técnico de um algoritmo, medido isoladamente por métricas como sensibilidade ou área sob a curva ROC, não garante benefício público quando os sistemas são implantados em redes assistenciais desiguais, com infraestrutura, formação e recursos distintos. A contribuição do artigo está em integrar requisitos técnicos, éticos, jurídicos e administrativos em um quadro aplicável ao longo de todo o ciclo de vida da inteligência artificial em saúde, da aquisição ao eventual descomissionamento, oferecendo subsídios tanto para gestores de serviços quanto para formuladores de políticas públicas.

Palavras-chave: inteligência artificial; saúde; governança; equidade; transparência; capacidade institucional.

ABSTRACT

The incorporation of artificial intelligence into healthcare systems raises governance challenges that go beyond the technical validation of algorithms. This article examines how normative principles can guide the responsible adoption of diagnostic, triage, and clinical management systems, with particular attention to equity in care, transparency of automated decisions, and the strengthening of institutional capacity within health organizations. The discussion draws on an integrative review of the literature on clinical governance, patient safety, distributive justice, algorithmic explainability, and organizational capabilities, engaging contributions published in journals such as Nature Medicine, The New England Journal of Medicine, Science, and BMC Medicine, as well as the World Health Organization guidance on ethics and governance of artificial intelligence for health. Building on this foundation, the article proposes a model composed of six interrelated dimensions: legitimate clinical purpose, data quality and representativeness, contextual validation, effective human oversight, contestation mechanisms, and continuous monitoring. It argues that the technical performance of an algorithm, measured in isolation through metrics such as sensitivity or the area under the ROC curve, does not guarantee public benefit when systems are deployed across unequal care networks with disparate infrastructure, training, and resources. The article's contribution lies in integrating technical, ethical, legal, and administrative requirements into a framework applicable across the entire lifecycle of artificial intelligence in healthcare, from procurement to eventual decommissioning, offering guidance for both service managers and policy makers.

Keywords: artificial intelligence; healthcare; governance; equity; transparency; institutional capacity.

RESUMEN

e-ISSN: 3086-7037

La incorporación de la inteligencia artificial en los sistemas de salud plantea desafíos de gobernanza que van más allá de la validación técnica de los algoritmos. Este artículo examina cómo los principios normativos pueden orientar la adopción responsable de sistemas de diagnóstico, triaje y gestión clínica, con especial atención a la equidad en la atención, la transparencia de las decisiones automatizadas y el fortalecimiento de la capacidad institucional de las organizaciones de salud. La discusión parte de una revisión integradora de la literatura sobre gobernanza clínica, seguridad del paciente, justicia distributiva, explicabilidad algorítmica y capacidades organizacionales, en diálogo con publicaciones como Nature Medicine, The New England Journal of Medicine, Science y BMC Medicine, además de las directrices de la Organización Mundial de la Salud sobre ética y gobernanza de la inteligencia artificial para la salud. A partir de esa base se propone un modelo con seis dimensiones interrelacionadas: finalidad clínica legítima, calidad y representatividad de los datos, validación contextual, supervisión humana efectiva, mecanismos de impugnación y monitoreo continuo. Se sostiene que el desempeño técnico de un algoritmo, medido de forma aislada mediante métricas como la sensibilidad o el área bajo la curva ROC, no garantiza beneficio público cuando los sistemas se implementan en redes asistenciales desiguales, con infraestructura, formación y recursos distintos. La contribución del artículo consiste en integrar requisitos técnicos, éticos, jurídicos y administrativos en un marco aplicable a todo el ciclo de vida de la inteligencia artificial en salud, desde la adquisición hasta el eventual retiro del sistema, ofreciendo insumos tanto para gestores de servicios como para formuladores de políticas públicas.

Palabras clave: inteligencia artificial; salud; gobernanza; equidad; transparencia; capacidad institucional.

INTRODUÇÃO

Sistemas de inteligência artificial já fazem parte da rotina de muitos serviços de saúde. Eles apoiam a leitura de exames de imagem, ajudam a estratificar risco cardiovascular, sinalizam pacientes com maior chance de deterioração clínica, organizam filas de espera e auxiliam no planejamento de recursos hospitalares. Topol (2019) descreveu esse movimento como parte de uma transição para uma medicina de alto desempenho, na qual algoritmos de aprendizado de máquina processam volumes de dados que ultrapassam a capacidade humana de análise manual. Essa mudança traz ganhos reais em determinados contextos, mas também introduz riscos que não se limitam a falhas técnicas do software.

O primeiro risco é o erro clínico decorrente de modelos mal calibrados ou aplicados fora do contexto para o qual foram validados. Um algoritmo treinado com dados de um hospital universitário de grande porte pode ter desempenho muito diferente quando usado em uma unidade básica de saúde com outro perfil de pacientes. Kelly et al. (2019), em revisão publicada na BMC Medicine, mostraram que a maioria dos estudos sobre inteligência artificial em saúde carece de validação externa robusta, o que limita a generalização dos resultados relatados em artigos técnicos para a prática clínica cotidiana.

O segundo risco envolve a opacidade dos modelos. Redes neurais profundas e outros métodos de aprendizado de máquina frequentemente funcionam como caixas-pretas, produzindo recomendações sem uma explicação intuitiva sobre o raciocínio que as gerou. Essa característica dificulta a supervisão clínica e a responsabilização quando algo dá errado, um problema já discutido por Verghese, Shah e Harrington (2018) ao alertarem que a confiança cega em recomendações computacionais pode enfraquecer o julgamento médico em vez de fortalecê-lo.

O terceiro risco, talvez o mais estudado nos últimos anos, é a desigualdade. Obermeyer et al. (2019), em artigo publicado na revista Science, demonstraram que um algoritmo amplamente utilizado nos Estados Unidos para identificar pacientes com necessidades de cuidado mais complexas discriminava sistematicamente pacientes negros, porque usava gastos em saúde como variável indireta (proxy) para necessidade clínica. Como pacientes negros historicamente têm menor acesso a serviços, mesmo com condições de saúde equivalentes, o modelo subestimava sua gravidade. O caso tornou-se referência obrigatória em qualquer discussão sobre governança de inteligência artificial em saúde, porque mostra que um sistema com boas métricas agregadas de desempenho pode, ainda assim, produzir dano distributivo grave.

Esses três riscos, erro clínico, opacidade e desigualdade, não são falhas de engenharia que se resolvem apenas com modelos mais sofisticados. Eles decorrem de decisões institucionais sobre como a tecnologia é escolhida, implantada, supervisionada e revisada. É esse o argumento central deste artigo: a governança da inteligência artificial em saúde precisa traduzir princípios éticos amplos em responsabilidades concretas, procedimentos verificáveis e critérios que permitam auditoria posterior. Sem essa tradução, decisões relevantes tendem a migrar para os fornecedores de tecnologia ou para especialistas isolados, reduzindo a capacidade da organização de saúde

de exercer supervisão efetiva sobre ferramentas que afetam diretamente a vida de pacientes.

Um segundo eixo do argumento trata da distribuição desigual de capacidades entre organizações de saúde. Hospitais universitários, redes privadas de grande porte e sistemas públicos bem financiados dispõem de equipes de ciência de dados, departamentos jurídicos e poder de negociação com fornecedores. Unidades menores, especialmente em regiões com menor investimento em saúde, geralmente não têm essa estrutura. Se a governança da inteligência artificial for tratada apenas como um problema técnico a ser resolvido internamente por cada organização, o resultado provável é a concentração de benefícios nas instituições já mais capacitadas, ampliando desigualdades entre territórios e populações. Políticas de apoio, padrões compartilhados e infraestrutura pública de avaliação podem atenuar essa assimetria, mas exigem decisão política deliberada.

Um terceiro eixo diz respeito à necessidade de monitoramento contínuo. O desempenho de um sistema de inteligência artificial não é estático. Mudanças na população atendida, em protocolos clínicos, em equipamentos de captura de imagem ou mesmo em hábitos de codificação de prontuário podem alterar substancialmente a qualidade das previsões ao longo do tempo, fenômeno conhecido na literatura técnica como desvio de dados ou data drift. Sendak et al. (2020), em estudo publicado na NPJ Digital Medicine sobre a implementação de um modelo preditivo de sepse em um sistema hospitalar americano, mostraram como o processo de validação, ajuste e monitoramento contínuo foi tão importante quanto o desenvolvimento inicial do algoritmo. Sem revisão periódica, um sistema validado em determinado momento pode se tornar inadequado sem que ninguém perceba, até que o dano já tenha ocorrido.

Por fim, cabe destacar a dimensão da prestação de contas. Quanto maior o potencial de um sistema afetar direitos, recursos ou oportunidades de tratamento, maior deve ser o nível de documentação, escrutínio e possibilidade de recurso por parte dos pacientes afetados. Essa proporcionalidade entre risco e controle é um princípio recorrente tanto na literatura sobre regulação de dispositivos médicos quanto nas diretrizes recentes sobre inteligência artificial, incluindo o documento da Organização Mundial da Saúde sobre ética e governança da inteligência artificial para a saúde (World Health Organization, 2021) e a abordagem baseada em risco adotada pelo Regulamento Europeu de Inteligência Artificial.

Este artigo está organizado da seguinte forma. Após esta introdução, apresenta-se o referencial teórico que sustenta a proposta, articulando governança clínica, segurança do paciente, justiça distributiva e teoria das capacidades organizacionais. Em seguida, desenvolve-se cada uma das seis dimensões do modelo proposto, com discussão sobre finalidade clínica, dados, validação, transparência, supervisão humana, contestação, monitoramento e capacidade institucional. Um bloco específico trata de contratação, regulação, formação profissional, interoperabilidade e avaliação econômica, temas frequentemente deixados de fora de discussões estritamente técnicas sobre inteligência artificial em saúde. O artigo encerra com uma discussão ampliada sobre as implicações institucionais do modelo, uma análise de suas limitações e considerações finais sobre agenda de pesquisa e políticas públicas.

REFERENCIAL TEÓRICO

Inteligência artificial como tecnologia sociotécnica

Um erro comum em discussões sobre inteligência artificial em saúde é tratar o algoritmo como um objeto isolado, cujo valor pode ser avaliado apenas por métricas de desempenho estatístico. A literatura sobre sistemas sociotécnicos, aplicada à informática em saúde desde os trabalhos clássicos de Coiera (2007) sobre sistemas de informação clínica, argumenta o contrário: qualquer tecnologia opera dentro de uma rede composta por pessoas, rotinas de trabalho, infraestrutura física e regras organizacionais. Um modelo de aprendizado de máquina com excelente desempenho em condições de teste pode falhar completamente quando inserido em um fluxo de trabalho que os profissionais não conseguem seguir, ou quando gera alertas em volume superior à capacidade de resposta da equipe.

Cabitza, Rasoini e Gensini (2017), em artigo publicado no JAMA sobre consequências não intencionais do aprendizado de máquina em medicina, chamaram atenção para o fenômeno da automação complacente, no qual profissionais passam a aceitar recomendações algorítmicas sem questionamento crítico, mesmo quando o contexto clínico sugere cautela. O oposto também ocorre: sistemas percebidos como pouco confiáveis são simplesmente ignorados, tornando o investimento tecnológico inútil. Ambos os desfechos resultam de má integração sociotécnica, não necessariamente de falha do modelo em si.

Essa perspectiva justifica por que a avaliação de um sistema de inteligência artificial em saúde não pode se limitar a testes de bancada. É preciso

considerar como profissionais interpretam recomendações, como o sistema se encaixa na carga de trabalho existente e como a organização responde quando o algoritmo erra. Esse é o pano de fundo teórico sobre o qual se apoiam as seis dimensões propostas mais adiante.

Governança clínica e segurança do paciente

A tradição de governança clínica, consolidada no Reino Unido a partir da década de 1990 como resposta a falhas assistenciais graves, oferece um vocabulário útil para pensar a inteligência artificial em saúde. Governança clínica pode ser entendida como o conjunto de estruturas por meio das quais organizações de saúde são responsáveis por melhorar continuamente a qualidade de seus serviços e salvaguardar padrões elevados de cuidado. Aplicada à inteligência artificial, essa tradição sugere que a adoção de um algoritmo deve seguir o mesmo rigor exigido para a introdução de um novo medicamento ou procedimento: evidência de eficácia, avaliação de segurança, treinamento da equipe e monitoramento pós-implantação.

Wiens et al. (2019), em artigo de referência publicado na Nature Medicine intitulado "Do no harm: a roadmap for responsible machine learning for health care", sistematizaram essa analogia ao propor um roteiro para desenvolvimento responsável de aprendizado de máquina em saúde, cobrindo desde a formulação do problema clínico até a manutenção pós-implantação. Os autores argumentam que a maior parte dos problemas éticos associados à inteligência artificial em saúde não decorre de má-fé, mas de lacunas em processos, como ausência de validação prospectiva, falta de monitoramento contínuo e escassez de mecanismos para reportar e corrigir falhas.

Justiça distributiva e vieses algorítmicos

A discussão sobre equidade em inteligência artificial em saúde dialoga diretamente com a tradição da justiça distributiva na bioética, que pergunta como benefícios e riscos de uma intervenção devem ser distribuídos entre diferentes grupos sociais. Modelos de aprendizado de máquina são treinados a partir de dados históricos, e dados históricos refletem as desigualdades de acesso, registro e qualidade de atendimento que já existem no sistema de saúde. Quando um grupo populacional aparece com menor frequência, menor qualidade de registro ou desfechos sistematicamente piores nos dados de treinamento, o modelo tende a reproduzir e, em alguns casos, amplificar essas diferenças.

O caso já mencionado de Obermeyer et al. (2019) ilustra esse mecanismo com clareza, mas está longe de ser isolado. Panch, Mattie e Atun (2019), em artigo sobre inteligência artificial e viés algorítmico publicado no Journal of Global Health, argumentam que vieses podem entrar no ciclo de desenvolvimento de um sistema de inteligência artificial em várias etapas, desde a escolha do problema a ser resolvido até a definição das variáveis de desfecho, passando pela composição da base de dados de treinamento e pelos critérios de validação. Por isso, os autores defendem que a auditoria de equidade não pode ser uma etapa isolada ao final do desenvolvimento, mas precisa acompanhar todo o processo.

Rajkomar, Hardt, Howell, Corrado e Chin (2018), em artigo publicado na Annals of Internal Medicine sobre justiça em aprendizado de máquina aplicado à saúde, propõem uma distinção útil entre diferentes noções técnicas de equidade, como paridade estatística, igualdade de oportunidade e calibração equilibrada entre subgrupos, mostrando que essas definições podem entrar em conflito entre si e que a escolha de um critério de equidade é, em última instância, uma decisão de valor, não apenas uma questão técnica. Essa constatação reforça o argumento deste artigo de que a governança de inteligência artificial em saúde exige deliberação explícita sobre valores, e não apenas otimização de métricas.

Explicabilidade e confiança

A capacidade de explicar por que um sistema chegou a determinada recomendação é frequentemente tratada como sinônimo de transparência, mas a literatura recente sugere maior cautela. Explicações técnicas detalhadas, como mapas de saliência em imagens ou coeficientes de importância de variáveis, podem ser incompreensíveis para profissionais sem formação em ciência de dados e, ainda mais, para pacientes leigos. Amann et al. (2020), em artigo sobre explicabilidade da inteligência artificial em saúde publicado no BMC Medical Informatics and Decision Making, argumentam que a explicabilidade precisa ser entendida em múltiplas dimensões, incluindo explicabilidade técnica, justificativa clínica e comunicação apropriada ao público, e que essas dimensões atendem a necessidades diferentes de diferentes atores.

Do ponto de vista da confiança institucional, a transparência cumpre uma função que vai além da explicação individual de cada previsão. Ela sustenta a possibilidade de auditoria externa, de comparação entre sistemas concorrentes e de responsabilização quando ocorre dano. Price e Cohen (2019), discutindo privacidade e big data médico na Nature Medicine,

também alertam que a transparência sobre como os dados dos pacientes são usados é condição prévia para a legitimidade de qualquer sistema de inteligência artificial em saúde, especialmente quando dados sensíveis circulam entre instituições públicas e privadas.

Capacidades organizacionais e teoria da capacidade estatal

Por fim, o modelo proposto neste artigo se apoia na literatura sobre capacidades organizacionais e capacidade estatal, que argumenta que a implementação efetiva de qualquer política ou tecnologia depende de recursos humanos qualificados, infraestrutura adequada, processos de coordenação e mecanismos de aprendizagem institucional. Sun e Medaglia (2019), em estudo sobre adoção de inteligência artificial em hospitais públicos chineses publicado na Government Information Quarterly, identificaram que barreiras organizacionais, como resistência de profissionais, falta de padronização de dados e ausência de liderança comprometida, foram tão relevantes quanto limitações técnicas para explicar o sucesso ou fracasso de projetos de inteligência artificial em saúde.

Essa literatura sustenta o argumento de que a governança responsável de inteligência artificial em saúde não pode ser reduzida a um checklist ético aplicado uma única vez. Ela exige capacidade contínua de avaliação, ajuste e decisão, distribuída de forma desigual entre organizações, o que torna a política pública de apoio institucional um componente inseparável da agenda ética e técnica.

FINALIDADE CLÍNICA E PROPORCIONALIDADE

O primeiro elemento do modelo proposto é a exigência de finalidade clínica legítima. Antes de discutir arquitetura de modelo ou qualidade de dados, uma organização precisa responder a uma pergunta simples: qual problema clínico ou organizacional o sistema pretende resolver, para qual população e com qual benefício esperado. Sistemas adotados sem essa definição clara tendem a ser avaliados apenas por sua sofisticação técnica, o que não corresponde necessariamente a utilidade clínica.

A proporcionalidade decorre diretamente dessa exigência. Um sistema de triagem que prioriza pacientes na fila de um pronto-socorro tem potencial de causar dano muito maior do que uma ferramenta que organiza agenda administrativa. Por isso, o nível de evidência exigido, a intensidade da supervisão humana e o rigor da documentação devem ser proporcionais ao risco potencial da aplicação, e não uniformes para toda e qualquer ferramenta rotulada como inteligência artificial. Essa lógica de graduação

por risco está presente tanto na regulamentação de dispositivos médicos baseados em software, como nas diretrizes da Food and Drug Administration para software como dispositivo médico, quanto na abordagem do Regulamento Europeu de Inteligência Artificial, que classifica sistemas usados em saúde predominantemente como de alto risco, sujeitos a requisitos reforçados de documentação, gestão de risco e supervisão humana.

Char, Shah e Magnus (2018), em artigo influente publicado no New England Journal of Medicine sobre implementação de aprendizado de máquina em saúde, alertaram para o risco de adoção de tecnologia por pressão competitiva ou modismo, sem que a organização tenha clareza sobre o problema a ser resolvido. Os autores defendem que decisões sobre adoção de inteligência artificial em saúde deveriam seguir processo semelhante ao de avaliação de novas terapias, com etapas de evidência, ensaio controlado quando possível e revisão por comitês independentes antes da adoção em larga escala.

Na prática institucional, essa exigência se traduz em documentos que registrem finalidade, população-alvo, alternativas consideradas, benefício esperado e critérios de sucesso, aprovados antes da aquisição do sistema. Esse registro cumpre função dupla: orienta a escolha entre fornecedores concorrentes e cria uma base de referência para avaliação posterior, permitindo comparar o que foi prometido com o que efetivamente ocorreu após a implantação.

DADOS, REPRESENTATIVIDADE E EQUIDADE

Modelos de aprendizado de máquina aprendem padrões a partir de dados históricos. Quando esses dados refletem desigualdades de acesso, registro ou qualidade de atendimento, o modelo tende a reproduzir tais desigualdades em suas recomendações. Esse problema, já discutido no referencial teórico a propósito do caso relatado por Obermeyer et al. (2019), exige que a governança institucional trate a composição da base de dados como questão central, e não como detalhe técnico a ser delegado inteiramente à equipe de desenvolvimento.

Uma prática recomendada, presente em diversas diretrizes recentes, é a avaliação de desempenho por subgrupos populacionais, como raça, gênero, idade, região geográfica e condição socioeconômica, além do desempenho agregado. Um sistema pode apresentar excelente acurácia geral e, ainda assim, ter desempenho significativamente pior em determinado subgrupo, o

que passaria despercebido em relatórios que apresentam apenas médias. Vayena, Blasimme e Cohen (2018), em artigo sobre desafios éticos do aprendizado de máquina em medicina publicado na PLOS Medicine, defendem que a transparência sobre a composição da base de treinamento, incluindo suas limitações de representatividade, deveria ser exigência mínima para qualquer sistema submetido à aprovação regulatória ou à aquisição institucional.

A governança deve também considerar o ciclo de retroalimentação entre uso do sistema e produção de novos dados. Se um algoritmo de triagem direciona menos recursos a determinado grupo, os desfechos desse grupo tendem a piorar, gerando dados que, se usados para retreinar o modelo no futuro, reforçam o padrão inicial de discriminação. Esse ciclo de retroalimentação é particularmente perigoso porque pode ocorrer de forma silenciosa, sem que gestores percebam a origem do problema. Documentar a origem dos dados, revisar periodicamente sua composição e manter processos de auditoria de equidade são medidas concretas para reduzir esse risco.

VALIDAÇÃO CLÍNICA CONTEXTUAL

Métricas de desempenho obtidas em ambiente de desenvolvimento, como sensibilidade, especificidade e área sob a curva ROC calculadas sobre uma base de teste retrospectiva, não substituem a validação no contexto real de uso. A diferença entre desempenho em condições controladas e desempenho em prática clínica cotidiana é um tema recorrente na literatura, muitas vezes chamada de lacuna entre validação técnica e utilidade clínica.

Kelly et al. (2019) revisaram estudos sobre inteligência artificial em saúde e constataram que grande parte deles não realiza validação externa, isto é, teste do modelo em uma população ou instituição diferente daquela usada no desenvolvimento. Sem essa etapa, não é possível saber se o desempenho relatado se mantém quando o sistema é transferido para outro contexto assistencial, com outro perfil de pacientes, outros equipamentos de captura de dados e outras rotinas de trabalho.

A validação contextual também precisa avaliar o impacto do sistema sobre decisões clínicas reais e sobre desfechos de pacientes, não apenas a acurácia isolada da previsão. Um modelo pode prever corretamente o risco de determinada complicação e, ainda assim, não gerar nenhum benefício, se a recomendação não altera a conduta clínica ou se os profissionais não confiam o suficiente na ferramenta para agir de acordo com ela. Estudos

prospectivos, comparando desfechos antes e depois da implantação, e estratégias de implantação gradual, começando por unidades piloto antes da expansão para toda a rede, são formas de reduzir o risco de generalizar prematuramente resultados obtidos em condições artificiais.

TRANSPARÊNCIA E EXPLICABILIDADE

Profissionais de saúde e pacientes precisam compreender, em algum grau, o papel que um sistema de inteligência artificial desempenha na decisão clínica. Isso não significa que cada usuário precise entender a arquitetura matemática do modelo, mas que a organização deve garantir explicações adequadas ao contexto e ao público, evitando tanto o excesso de tecnicismo quanto simplificações que induzam a confiança excessiva ou a rejeição infundada.

Amann et al. (2020) propõem distinguir entre diferentes camadas de explicabilidade: a explicação técnica voltada a desenvolvedores e auditores, a justificativa clínica voltada a profissionais de saúde, que precisam entender por que determinada recomendação faz sentido diante do quadro do paciente, e a comunicação voltada ao paciente, que exige linguagem acessível e atenção a diferenças de letramento em saúde. A ausência de qualquer uma dessas camadas compromete a governança do sistema, ainda que as demais estejam bem desenvolvidas.

A documentação institucional cumpre papel central nessa dimensão. Fichas técnicas que descrevam a finalidade do sistema, a base de dados utilizada, as limitações conhecidas, os subgrupos populacionais em que o desempenho foi avaliado e o histórico de atualizações deveriam estar disponíveis para consulta por profissionais e, quando pertinente, por pacientes e órgãos reguladores. Esse tipo de documentação, análogo à bula de um medicamento, transforma a transparência de princípio abstrato em prática institucional verificável.

SUPERVISÃO HUMANA

A presença de um profissional entre o sistema e a decisão final é frequentemente apresentada como garantia suficiente de segurança, mas a literatura mostra que essa suposição é frágil. Supervisão humana efetiva exige que o profissional tenha tempo, autonomia e competência para revisar criticamente a recomendação, e não apenas para referendá-la formalmente. Quando a carga de trabalho é excessiva, quando a organização trata divergências em relação ao sistema como desvio a ser penalizado, ou quando

o resultado algorítmico é apresentado com uma autoridade que desestimula questionamento, a supervisão se torna simbólica.

Cabitza et al. (2017) descrevem esse fenômeno como uma forma de automação complacente, em que a presença humana no processo deixa de funcionar como camada de segurança e passa a legitimar decisões que, na prática, não foram de fato revisadas. Protocolos institucionais precisam definir explicitamente em quais situações a recomendação algorítmica pode ser aceita diretamente, em quais exige revisão obrigatória e em quais deve ser desconsiderada diante de sinais clínicos contraditórios. Esses protocolos também devem prever registro da decisão do profissional, permitindo posteriormente avaliar se a supervisão humana está de fato influenciando os desfechos ou funcionando apenas como formalidade.

A formação continuada é condição necessária para que a supervisão humana seja substantiva. Profissionais precisam compreender, em termos gerais, como o sistema funciona, quais são suas limitações conhecidas e em que situações seu desempenho tende a ser menos confiável. Sem esse conhecimento mínimo, a autonomia formal de discordar da recomendação algorítmica não se traduz em capacidade real de exercê-la.

CONTESTAÇÃO E PARTICIPAÇÃO DO PACIENTE

Pacientes afetados por decisões apoiadas em sistemas automatizados precisam ter acesso a informação sobre essa participação e a canais efetivos de questionamento. Esse direito de contestação cumpre função dupla: protege interesses individuais dos pacientes e funciona como mecanismo de correção para a própria organização, já que reclamações e questionamentos frequentemente revelam falhas que não aparecem em métricas agregadas de desempenho.

A comunicação sobre o uso de inteligência artificial precisa considerar diferenças de letramento em saúde e de familiaridade com tecnologia entre diferentes grupos de pacientes. Uma explicação padronizada, escrita em linguagem técnica, pode cumprir requisitos formais de transparência sem de fato informar o paciente sobre o que está em jogo. Iniciativas de participação social, como conselhos de usuários e consultas públicas sobre novos sistemas antes de sua adoção em larga escala, têm sido recomendadas por diretrizes internacionais como forma de incorporar essa perspectiva de forma estruturada, e não apenas reativa.

A Organização Mundial da Saúde (2021), em seu documento sobre ética e governança da inteligência artificial para a saúde, destaca a participação

pública como um dos princípios centrais para a adoção responsável dessas tecnologias, argumentando que decisões sobre sistemas que afetam diretamente o cuidado em saúde não deveriam ser tomadas exclusivamente por desenvolvedores de tecnologia e gestores, sem envolvimento das comunidades e pacientes afetados.

MONITORAMENTO E GESTÃO DE INCIDENTES

O desempenho de um sistema de inteligência artificial pode se deteriorar depois da implantação, mesmo sem qualquer alteração no código do modelo. Mudanças na população atendida, em protocolos clínicos, em equipamentos usados para captura de exames ou em práticas de registro de prontuário podem alterar a distribuição dos dados de entrada, um fenômeno que a literatura técnica chama de desvio de dados. Sem monitoramento contínuo, esse tipo de deterioração pode passar despercebido por longos períodos.

Sendak et al. (2020) descrevem, a partir da experiência de implementação de um modelo de previsão de sepse, como a etapa de monitoramento pós-implantação exigiu tanta atenção institucional quanto o próprio desenvolvimento do algoritmo, incluindo ajustes de limiar de alerta, revisão periódica de desempenho e canais para que profissionais reportassem casos em que a recomendação pareceu inadequada. Esse tipo de estrutura de acompanhamento deveria ser parte obrigatória de qualquer projeto de inteligência artificial em saúde, não um complemento opcional.

A gestão de incidentes exige registro sistemático de eventos adversos relacionados ao uso do sistema, investigação das causas e comunicação transparente sobre medidas corretivas. Quando o risco identificado for grave e não puder ser mitigado rapidamente, a suspensão temporária do uso do sistema deve ser uma opção institucional disponível e efetivamente exercida, e não apenas uma possibilidade teórica prevista em contrato.

CAPACIDADE INSTITUCIONAL

A governança responsável de inteligência artificial em saúde depende de capacidade institucional concreta: equipes multidisciplinares capazes de avaliar propostas técnicas, infraestrutura de dados adequada, contratos bem redigidos e liderança comprometida com a supervisão contínua dos sistemas adotados. Organizações precisam desenvolver competência para avaliar fornecedores de forma crítica, o que inclui capacidade de solicitar e interpretar evidências de validação, negociar cláusulas contratuais sobre auditoria e desempenho, e formar profissionais para uso adequado da tecnologia.

Essa capacidade não é distribuída igualmente entre organizações de saúde. Redes com poucos recursos, especialmente em regiões de menor desenvolvimento econômico, dependem quase inteiramente de fornecedores externos para avaliar e ajustar sistemas de inteligência artificial, o que reduz sua autonomia decisória e sua capacidade de negociação. Sun e Medaglia (2019) identificaram, em seu estudo sobre adoção de inteligência artificial em hospitais chineses, que barreiras organizacionais, e não apenas limitações técnicas, foram determinantes para explicar o sucesso ou fracasso de diferentes iniciativas, reforçando a ideia de que investimento em capacidade institucional é tão importante quanto investimento em tecnologia propriamente dita.

Políticas públicas de apoio, incluindo padrões técnicos compartilhados, infraestrutura pública de avaliação independente e programas de formação, podem reduzir essa assimetria e evitar que a inteligência artificial em saúde se torne mais um fator de concentração de vantagens entre instituições já bem capacitadas.

QUADRO PROPOSTO

O quadro desenvolvido neste artigo integra as seis dimensões discutidas: finalidade clínica legítima, qualidade e representatividade dos dados, validação contextual, transparência e explicabilidade, supervisão humana efetiva e mecanismos de contestação, além do monitoramento contínuo que atravessa todas as demais. Essas dimensões não devem ser entendidas como etapas sequenciais e isoladas, mas como aspectos que se retroalimentam ao longo de todo o ciclo de vida do sistema, da aquisição ao eventual descomissionamento.

A decisão de adotar um sistema de inteligência artificial em saúde deve ser tratada como revisável desde o início, e não como um compromisso definitivo assumido no momento da aquisição. Isso significa manter registros que permitam comparar o desempenho previsto com o desempenho observado, revisar periodicamente a pertinência da tecnologia diante de alternativas disponíveis e, quando necessário, interromper seu uso. Essa orientação para valor público, mais do que para eficiência técnica isolada, é o fio condutor que une as seis dimensões do modelo.

CONTRATAÇÃO, REGULAÇÃO E RESPONSABILIDADE

A aquisição de sistemas de inteligência artificial em saúde exige critérios que combinem avaliação clínica, técnica e jurídica. Editais e contratos devem especificar requisitos de desempenho mínimo, documentação técnica

exigida, padrões de segurança da informação, requisitos de interoperabilidade com sistemas já existentes e condições de suporte técnico continuado. Sem esses elementos, a organização compradora fica em posição de desvantagem informacional em relação ao fornecedor, incapaz de avaliar adequadamente se o produto entregue corresponde ao que foi prometido.

A definição de responsabilidades entre desenvolvedores de tecnologia, instituições de saúde e profissionais que utilizam o sistema precisa ser explícita. A multiplicação de atores envolvidos no desenvolvimento e uso de um sistema de inteligência artificial não pode resultar em lacunas de responsabilização, situação em que, diante de um dano, cada parte alega que a falha pertence a outra. Protocolos de investigação de incidentes, definidos previamente e não improvisados após a ocorrência de um problema, ajudam a esclarecer causas e atribuir responsabilidades de forma justa.

Regulação externa e governança institucional interna são complementares, não substitutas uma da outra. A conformidade formal com exigências regulatórias, como registro em agência sanitária ou certificação de qualidade, não elimina a necessidade de avaliação local sobre adequação do sistema ao contexto específico de uso. Organizações precisam acompanhar atualizações regulatórias e mudanças na forma como o sistema é utilizado na prática, que podem se afastar do uso originalmente previsto e aprovado.

Contratos também devem prever mecanismos de auditoria independente, portabilidade de dados em caso de troca de fornecedor e condições claras para encerramento da relação contratual. A dependência excessiva de um único fornecedor, sem cláusulas que garantam continuidade e transparência, compromete a capacidade da instituição de exercer controle efetivo sobre a tecnologia adotada.

FORMAÇÃO PROFISSIONAL E CULTURA DE SEGURANÇA

Profissionais de saúde precisam compreender, ainda que em nível não técnico, as capacidades e os limites dos sistemas de inteligência artificial com os quais trabalham. Programas de formação deveriam incluir noções básicas sobre como interpretar resultados algorítmicos, quais tipos de viés podem afetar as recomendações, princípios de privacidade de dados clínicos e estratégias de comunicação com pacientes sobre o papel da tecnologia no cuidado recebido. O objetivo não é transformar cada profissional em

especialista em ciência de dados, mas desenvolver julgamento crítico suficiente para usar a ferramenta de forma segura.

Uma cultura organizacional que permita relatar erros e divergências sem medo de punição é condição necessária para que o monitoramento contínuo funcione na prática. Ambientes punitivos tendem a reduzir o relato espontâneo de problemas, escondendo falhas até que se tornem graves. Canais confidenciais de relato e análise sistêmica dos incidentes reportados, em vez de responsabilização individual imediata, fortalecem a capacidade de aprendizagem da organização.

Equipes multidisciplinares, que reúnam conhecimento clínico, técnico e ético, ajudam a identificar problemas que passariam despercebidos em análises puramente técnicas. Reuniões periódicas de revisão de desempenho, com participação de diferentes categorias profissionais, permitem acompanhar tanto indicadores quantitativos quanto relatos qualitativos sobre a experiência de uso do sistema.

Pacientes e comunidades também precisam de informação acessível sobre como sistemas de inteligência artificial são usados em seu cuidado. Materiais de comunicação elaborados em linguagem simples, testados com usuários reais antes de sua divulgação, ajudam a construir compreensão sobre usos e direitos, contribuindo para a confiança institucional de forma mais consistente do que declarações genéricas sobre inovação tecnológica.

INTEROPERABILIDADE E CONTINUIDADE DO CUIDADO

Sistemas de inteligência artificial dependem de dados que frequentemente circulam entre diferentes serviços e níveis de atenção. A interoperabilidade técnica, que garante que sistemas diferentes consigam trocar dados, e a interoperabilidade semântica, que garante que esses dados mantenham o mesmo significado clínico quando transferidos entre sistemas, são condições essenciais para evitar fragmentação da informação e, com ela, perda de qualidade nas recomendações geradas.

A continuidade do cuidado pode ser prejudicada quando recomendações geradas por um sistema em determinado serviço não acompanham o paciente ao longo de sua trajetória assistencial. Um alerta de risco gerado durante uma internação, por exemplo, perde utilidade se não for comunicado à equipe de atenção primária responsável pelo acompanhamento após a alta. A integração entre sistemas de inteligência artificial e registros eletrônicos de saúde precisa ser planejada desde o início do projeto, e não tratada como ajuste posterior.

Trocas de fornecedor ou de versão de um sistema exigem processos de migração de dados que preservem histórico clínico relevante. Contratos de aquisição deveriam prever explicitamente condições de portabilidade e de preservação de registros, evitando que a mudança de tecnologia resulte em perda de informação assistencial importante. A interoperabilidade bem estruturada também facilita processos de auditoria externa e de pesquisa científica sobre o desempenho real dos sistemas em uso, desde que sejam respeitadas exigências de privacidade e de finalidade legítima no uso secundário dos dados.

AVALIAÇÃO ECONÔMICA E VALOR PÚBLICO

A decisão de implantar um sistema de inteligência artificial em saúde deve considerar seu custo total, que inclui não apenas a licença ou aquisição inicial, mas também infraestrutura de dados, formação de profissionais, manutenção técnica continuada e monitoramento de desempenho ao longo do tempo. Projeções de economia de custos apresentadas por fornecedores frequentemente não se concretizam quando processos manuais paralelos permanecem em funcionamento por precaução, ou quando a implantação exige investimentos complementares não previstos inicialmente.

A avaliação econômica de sistemas de inteligência artificial em saúde precisa incorporar dimensões de segurança, equidade e qualidade dos desfechos, não apenas indicadores de eficiência operacional. Redução no tempo de atendimento, por exemplo, não representa benefício real quando é obtida às custas de maior taxa de erro diagnóstico ou de barreiras adicionais de acesso para determinados grupos de pacientes.

Projetos-piloto, conduzidos em escala reduzida antes da expansão para toda a rede assistencial, ajudam a estimar custos reais e efeitos clínicos e organizacionais de forma mais confiável do que projeções feitas exclusivamente com base em dados fornecidos pelo desenvolvedor da tecnologia. Critérios objetivos para decidir sobre a ampliação de escala, definidos antes do início do piloto, evitam que decisões de expansão sejam tomadas por pressão comercial ou por entusiasmo tecnológico não fundamentado em evidência local.

A transparência sobre custos e resultados de projetos de inteligência artificial em saúde permite controle social e priorização adequada de recursos públicos. Como recursos destinados à saúde envolvem escolhas com custo de oportunidade elevado, investimentos em tecnologia precisam

ser justificados publicamente com a mesma exigência aplicada a outras decisões de alocação de recursos assistenciais.

AGENDA DE PESQUISA E POLÍTICAS

Pesquisas futuras sobre governança de inteligência artificial em saúde deveriam priorizar estudos conduzidos em ambientes reais de prática clínica, com populações diversas, em vez de depender exclusivamente de avaliações retrospectivas em bases de dados controladas. Estudos prospectivos, que acompanhem desfechos clínicos antes e depois da implantação de um sistema, oferecem evidência mais robusta sobre benefício real do que comparações puramente técnicas entre modelos.

É necessário aprofundar a compreensão sobre como recomendações algorítmicas interagem com o julgamento clínico profissional em condições reais de trabalho, incluindo fatores organizacionais como carga de trabalho, cultura institucional e relação de confiança entre equipe e tecnologia. Esses fatores frequentemente explicam variações de desempenho e de adesão que não podem ser previstas apenas a partir de características técnicas do modelo.

Políticas públicas têm papel importante no apoio a padrões técnicos comuns, infraestrutura compartilhada de avaliação independente e mecanismos de certificação que reduzam a assimetria de informação entre fornecedores e organizações de saúde, especialmente as de menor porte. Redes menores, com recursos limitados, precisam de apoio estruturado para participar da adoção responsável dessas tecnologias sem depender inteiramente da boa vontade de fornecedores comerciais.

Por fim, a governança de inteligência artificial em saúde precisa ser entendida como processo em evolução constante, que acompanha tanto o avanço tecnológico quanto o acúmulo de evidência sobre seus efeitos reais na prática clínica. Revisões periódicas de diretrizes institucionais e regulatórias são necessárias para manter o quadro de governança atualizado diante de mudanças que ocorrem em ritmo mais rápido do que os ciclos habituais de revisão normativa.

DISCUSSÃO AMPLIADA E IMPLICAÇÕES INSTITUCIONAIS

A análise desenvolvida ao longo deste artigo indica que mudanças tecnológicas produzem resultados consistentes apenas quando acompanhadas de desenho institucional adequado. A disponibilidade de ferramentas, dados e competências individuais não substitui regras de

coordenação, definição de autoridade decisória e mecanismos de prestação de contas. Organizações de saúde precisam explicitar finalidades, atribuir responsabilidades e definir critérios verificáveis de avaliação. Essa estrutura não elimina incerteza, mas cria condições para que decisões possam ser justificadas, revisadas e aprimoradas ao longo do tempo, em vez de permanecerem implícitas em escolhas informais de gestores ou fornecedores.

Um segundo aspecto relevante envolve a relação entre capacidade técnica e legitimidade institucional. Projetos de inteligência artificial em saúde podem apresentar bom desempenho operacional, medido por métricas técnicas, e ainda assim enfrentar resistência de profissionais e desconfiança de pacientes quando os afetados não compreendem seus objetivos ou não dispõem de canais reais de participação nas decisões sobre sua adoção. A legitimidade se constrói pela combinação entre evidência técnica, transparência processual e tratamento justo dos diferentes grupos afetados. Processos deliberativos não eliminam conflitos de interesse ou de valores, mas ajudam a tornar visíveis prioridades e consequências que análises exclusivamente técnicas tendem a ignorar.

A distribuição desigual de recursos entre organizações de saúde condiciona diretamente a governança dessas tecnologias. Instituições de maior porte contam com equipes especializadas, infraestrutura robusta e maior poder de negociação com fornecedores, enquanto organizações menores dependem quase inteiramente de apoio externo. Políticas de cooperação entre instituições e de compartilhamento de capacidades de avaliação podem reduzir essa diferença. Qualquer avaliação de impacto de um sistema de inteligência artificial em saúde deveria observar não apenas os resultados agregados, mas também se a inovação amplia ou reduz assimetrias entre territórios, grupos populacionais e diferentes unidades de uma mesma rede assistencial.

A proporcionalidade entre risco e controle, discutida na seção sobre finalidade clínica, tem implicações práticas importantes para o desenho de processos institucionais. Aplicações de baixo impacto, como ferramentas administrativas de agendamento, podem ser geridas com procedimentos relativamente simples. Decisões que afetam diretamente diagnóstico, tratamento ou priorização de acesso a recursos escassos exigem documentação detalhada, revisão por instâncias independentes e mecanismos claros de recurso para pacientes que se sintam prejudicados. Um sistema de governança que trate todas as aplicações com o mesmo grau

de rigor tende a produzir dois problemas opostos: burocracia excessiva para casos de baixo risco e controle insuficiente para casos de alto impacto.

A capacidade de aprendizagem institucional constitui outro componente decisivo para a sustentabilidade da governança proposta. Projetos de inteligência artificial em saúde deveriam ser tratados como processos adaptativos, formulados com hipóteses explícitas, indicadores de acompanhamento e momentos formais de revisão, em vez de implementações definitivas avaliadas apenas no lançamento. Resultados inesperados, sejam eles positivos ou negativos, precisam gerar ajustes concretos em modelos, contratos e rotinas de trabalho. Organizações que ocultam falhas, por receio de responsabilização ou de dano reputacional, perdem oportunidades relevantes de correção e tendem a repetir os mesmos problemas em projetos futuros. Ambientes que combinam responsabilização com espaço legítimo para correção de rota favorecem esse tipo de aprendizagem.

Do ponto de vista gerencial, a governança de inteligência artificial em saúde exige equipes genuinamente multidisciplinares, nas quais conhecimento técnico dialoga com experiência clínica operacional, formação jurídica, reflexão ética e capacidade de comunicação com pacientes e comunidades. Essa diversidade de competências reduz pontos cegos que costumam aparecer quando decisões ficam concentradas em um único grupo profissional, seja ele técnico, clínico ou administrativo. Ao mesmo tempo, equipes multidisciplinares precisam de linguagem comum e de processos de decisão bem definidos, sob risco de fragmentação e de decisões contraditórias entre diferentes áreas da mesma organização.

A sustentabilidade financeira e organizacional de projetos de inteligência artificial em saúde depende de recursos recorrentes, não apenas de investimento inicial de implantação. Iniciativas financiadas exclusivamente para o momento de lançamento tendem a se deteriorar quando termina o período de projeto ou de financiamento externo. Custos de manutenção, atualização de modelos, formação continuada de profissionais e monitoramento de desempenho precisam ser incluídos desde o planejamento orçamentário inicial. A decisão de continuar, ajustar ou encerrar uma iniciativa de inteligência artificial em saúde deve considerar de forma explícita o valor efetivamente produzido, os riscos identificados ao longo do uso e as alternativas disponíveis, inclusive a possibilidade de retorno a processos não automatizados quando estes se mostrarem mais adequados ao contexto.

Por fim, o desenvolvimento de sistemas de inteligência artificial em saúde deve permanecer orientado por finalidade pública e organizacional, e não pela sofisticação tecnológica em si. O valor de qualquer ferramenta decorre de sua capacidade de resolver problemas concretos, fortalecer a capacidade das instituições de saúde e produzir benefícios distribuídos de forma justa entre a população atendida. A avaliação de qualquer projeto deveria retornar periodicamente à pergunta que orientou sua criação: quais necessidades foram efetivamente atendidas, para quem, e com quais consequências para os diferentes grupos afetados.

LIMITAÇÕES DO MODELO E CRITÉRIOS PARA APLICAÇÃO

O modelo apresentado neste artigo tem natureza teórica e precisa ser adaptado ao contexto específico de cada organização. Serviços de saúde variam substancialmente em porte, missão institucional, disponibilidade de recursos, ambiente regulatório e maturidade tecnológica. A aplicação mecânica das recomendações aqui discutidas, sem essa adaptação, pode gerar novos problemas em vez de resolvê-los. Antes de implementar qualquer uma das dimensões propostas, é necessário mapear processos já existentes, identificar partes interessadas relevantes, avaliar riscos específicos do contexto local e reconhecer honestamente as capacidades disponíveis.

Outra limitação decorre da velocidade das mudanças tecnológicas no campo da inteligência artificial. Ferramentas, padrões técnicos e configurações de mercado evoluem rapidamente, enquanto regras institucionais e processos regulatórios costumam exigir mais tempo para se consolidar. A governança precisa ser estável o suficiente para oferecer segurança jurídica e operacional, mas flexível o bastante para incorporar novas evidências sem que isso exija reformulação completa a cada avanço tecnológico. Revisões programadas em intervalos regulares, e não apenas em resposta a crises, ajudam a equilibrar essas duas necessidades aparentemente contraditórias.

A disponibilidade e a qualidade dos dados usados para avaliação e monitoramento também limitam, na prática, a aplicação do modelo. Indicadores de desempenho podem estar incompletos, apresentar inconsistências entre diferentes fontes ou serem produzidos diretamente pelos próprios fornecedores da tecnologia, o que compromete sua independência. A triangulação entre diferentes fontes de evidência, incluindo auditorias externas e observações qualitativas do uso cotidiano do sistema, reduz a dependência de uma única medida de desempenho. Quando o grau de incerteza sobre o real funcionamento de um sistema for elevado,

decisões de expansão de uso deveriam ser graduais e reversíveis, evitando compromissos institucionais difíceis de desfazer.

Conflitos de interesse merecem atenção específica em qualquer aplicação prática do modelo. Fornecedores de tecnologia podem influenciar, direta ou indiretamente, a definição de critérios de sucesso e apresentar resultados de forma seletiva, destacando cenários favoráveis e omitindo limitações relevantes. Consultores contratados, pesquisadores vinculados a projetos e gestores em busca de resultados rápidos também podem ter incentivos que não coincidem plenamente com o interesse dos pacientes. Declarações formais de conflito de interesse, transparência sobre termos contratuais e revisão por partes independentes ajudam a mitigar, ainda que não eliminem completamente, esse tipo de viés institucional.

A aplicação efetiva do modelo exige liderança institucional comprometida. Sem apoio explícito da alta gestão, comitês de avaliação e protocolos de supervisão tendem a se tornar formalidades burocráticas com pouco efeito prático sobre as decisões reais da organização. Lideranças precisam disponibilizar recursos suficientes, responder de forma substantiva às recomendações produzidas por esses processos e aceitar, quando necessário, a possibilidade de interromper iniciativas em andamento, mesmo diante de investimento já realizado.

Os critérios de aplicação prática do modelo deveriam incluir relevância clínica ou organizacional da tecnologia proposta, proporcionalidade entre risco e nível de controle exigido, capacidade institucional disponível, participação adequada de profissionais e pacientes afetados, e estrutura de monitoramento contínuo. Cada um desses critérios pode ser registrado em uma matriz de decisão simples, que facilite a comparação entre alternativas tecnológicas concorrentes e mantenha registro das justificativas usadas em cada escolha, criando histórico útil para avaliações futuras.

Cabe destacar, por fim, que a decisão de não adotar determinada tecnologia, ou de encerrar seu uso após período de avaliação, constitui resultado legítimo e às vezes desejável do processo de governança. Inovação responsável em saúde inclui a capacidade institucional de reconhecer quando os benefícios esperados de um sistema não compensam seus custos e riscos reais. Processos de encerramento bem planejados devem preservar dados clínicos relevantes, garantir continuidade dos serviços prestados aos pacientes e manter registro do conhecimento acumulado durante o período de uso, de modo que essa experiência possa informar decisões futuras. Essas limitações não reduzem a utilidade do quadro proposto neste artigo;

ao contrário, indicam as condições necessárias para seu uso prudente. A principal contribuição do modelo está em organizar, de forma integrada, perguntas que costumam ser tratadas de maneira fragmentada por diferentes áreas de uma mesma organização de saúde.

CONSIDERAÇÕES FINAIS

A inteligência artificial pode contribuir de forma relevante para sistemas de saúde, apoiando diagnóstico, triagem, gestão de recursos e planejamento assistencial. Ao longo deste artigo, procurou-se argumentar que esses benefícios dependem de estruturas de governança capazes de traduzir princípios éticos amplos em responsabilidades concretas, procedimentos verificáveis e mecanismos efetivos de correção. Desempenho técnico isolado, medido por métricas estatísticas de acurácia ou discriminação, não garante equidade no atendimento, segurança para os pacientes nem fortalecimento da capacidade institucional das organizações de saúde envolvidas.

O modelo proposto, organizado em torno de seis dimensões inter-relacionadas, finalidade clínica legítima, qualidade e representatividade dos dados, validação contextual, transparência e explicabilidade, supervisão humana efetiva, mecanismos de contestação e monitoramento contínuo, busca oferecer diretrizes aplicáveis tanto a gestores de serviços de saúde quanto a formuladores de políticas públicas responsáveis pela regulação dessas tecnologias. A contribuição central do artigo está em tratar esses requisitos como componentes de um mesmo sistema, e não como listas de verificação independentes a serem cumpridas isoladamente.

Reconhece-se, contudo, que o modelo tem natureza teórica e que sua aplicação prática exige adaptação cuidadosa a cada contexto institucional específico. Pesquisas futuras poderiam testar a aplicabilidade do quadro em diferentes redes assistenciais, avaliando de forma empírica se organizações que adotam essas seis dimensões apresentam, de fato, melhores resultados de segurança, equidade e legitimidade institucional do que organizações que tratam a inteligência artificial predominantemente como questão técnica. Estudos comparativos entre sistemas de saúde com diferentes níveis de capacidade institucional e diferentes marcos regulatórios também poderiam esclarecer em que medida os princípios discutidos neste artigo se sustentam em contextos distintos daquele que motivou originalmente sua formulação.

Por fim, cabe reafirmar que a incorporação responsável da inteligência artificial aos sistemas de saúde não é apenas um desafio técnico a ser resolvido por engenheiros e cientistas de dados, nem apenas uma questão

regulatória a ser definida por legisladores. Trata-se de um processo contínuo de construção institucional, que exige coordenação entre diferentes atores, deliberação explícita sobre valores em disputa e disposição para revisar decisões diante de novas evidências. A governança discutida ao longo deste artigo representa uma tentativa de organizar essa complexidade em um quadro coerente, sem a pretensão de esgotar um debate que, dada a velocidade das transformações tecnológicas em curso, continuará em desenvolvimento por muitos anos.

Uma questão adicional, pouco explorada na literatura brasileira sobre o tema, diz respeito à relação entre governança de inteligência artificial e reforma mais ampla dos sistemas de saúde. Em contextos de sistemas públicos universais, como o Sistema Único de Saúde, decisões sobre adoção de tecnologia não podem ser tomadas isoladamente por cada unidade assistencial, sob risco de produzir um mosaico de práticas incompatíveis entre si, com padrões de qualidade e segurança muito diferentes de uma região para outra. A coordenação entre níveis de gestão, municipal, estadual e federal, torna-se necessária para garantir que princípios mínimos de governança sejam observados em toda a rede, sem impedir, contudo, adaptações locais justificadas por diferenças reais de contexto assistencial.

Essa tensão entre padronização nacional e adaptação local aparece de forma recorrente em outras áreas da política de saúde e não é exclusiva da inteligência artificial. A experiência acumulada com protocolos clínicos, sistemas de informação em saúde e políticas de regulação de acesso sugere que soluções bem-sucedidas costumam combinar diretrizes gerais definidas em nível central com margem de ajuste operacional em nível local. Aplicada à inteligência artificial, essa lógica sugere que agências regulatórias e órgãos de gestão do sistema de saúde deveriam definir requisitos mínimos de segurança, transparência e equidade, deixando para cada organização a tarefa de traduzir esses requisitos em processos compatíveis com sua realidade operacional.

Vale ainda considerar o papel da pesquisa acadêmica na sustentação desse processo de governança. Universidades e centros de pesquisa podem contribuir de forma relevante, não apenas desenvolvendo novos algoritmos, mas também produzindo avaliação independente sobre o desempenho real de sistemas já em uso, algo que fornecedores comerciais raramente têm incentivo para fazer de forma isenta. Parcerias entre instituições de pesquisa e serviços de saúde, com salvaguardas adequadas de proteção de dados e de conflito de interesse, podem funcionar como mecanismo

complementar de controle de qualidade, especialmente em sistemas de saúde com capacidade regulatória limitada.

Merece registro, por fim, a dimensão temporal da governança discutida neste artigo. Muitas das falhas relatadas na literatura sobre inteligência artificial em saúde não ocorreram no momento da implantação inicial, mas surgiram meses ou anos depois, à medida que o contexto de uso se afastava gradualmente das condições originais de validação. Isso reforça a ideia de que governança não é um evento único, localizado no início de um projeto, mas um processo contínuo que acompanha toda a vida útil do sistema. Organizações que tratam a aprovação inicial como ponto final do processo de avaliação tendem a ser surpreendidas por problemas que poderiam ter sido identificados e corrigidos a tempo, caso existisse estrutura permanente de acompanhamento.

REFERÊNCIAS

Amann, J., Blasimme, A., Vayena, E., Frey, D., & Madai, V. I. (2020). Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC Medical Informatics and Decision Making, 20(1), 310.

Cabitza, F., Rasoini, R., & Gensini, G. F. (2017). Unintended consequences of machine learning in medicine. JAMA, 318(6), 517-518.

Char, D. S., Shah, N. H., & Magnus, D. (2018). Implementing machine learning in health care: addressing ethical challenges. New England Journal of Medicine, 378(11), 981-983.

Coiera, E. (2007). Putting the technical back into socio-technical systems research. International Journal of Medical Informatics, 76, S98-S103.

Esteva, A., Chou, K., Yeung, S., Naik, N., Madani, A., Mottaghi, A., Liu, Y., Topol, E., Dean, J., & Socher, R. (2019). A guide to deep learning in healthcare. Nature Medicine, 25(1), 24-29.

Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G., & King, D. (2019). Key challenges for delivering clinical impact with artificial intelligence. BMC Medicine, 17(1), 195.

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447-453.

Panch, T., Mattie, H., & Atun, R. (2019). Artificial intelligence and algorithmic bias: implications for health systems. Journal of Global Health, 9(2), 010318.

Price, W. N., & Cohen, I. G. (2019). Privacy in the age of medical big data. Nature Medicine, 25(1), 37-43.

Rajkomar, A., Hardt, M., Howell, M. D., Corrado, G., & Chin, M. H. (2018). Ensuring fairness in machine learning to advance health equity. Annals of Internal Medicine, 169(12), 866-872.

Sendak, M. P., Ratliff, W., Sarro, D., Alderton, E., Futoma, J., Gao, M., Nichols, M., Revoir, M., Yashar, F., Miller, C., Kester, K., Sandhu, S., Corey, K., Brajer, N., Tan, C., Lin, A., Brown, T., Engelbosch, S., Anstrom, K., ... Balu, S. (2020). Real-world integration of a sepsis deep learning technology into routine clinical care: implementation study. NPJ Digital Medicine, 3(1), 84.

Sun, T. Q., & Medaglia, R. (2019). Mapping the challenges of artificial intelligence in the public sector: evidence from public healthcare. Government Information Quarterly, 36(2), 368-383.

Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. Nature Medicine, 25(1), 44-56.

Vayena, E., Blasimme, A., & Cohen, I. G. (2018). Machine learning in medicine: addressing ethical challenges. PLOS Medicine, 15(11), e1002689.

Verghese, A., Shah, N. H., & Harrington, R. A. (2018). What this computer needs is a physician: humanism and artificial intelligence. JAMA, 319(1), 19-20.

Wiens, J., Saria, S., Sendak, M., Ghassemi, M., Liu, V. X., Doshi-Velez, F., Jung, K., Heller, K., Kale, D., Saeed, M., Ossorio, P. N., Thadaney-Israni, S., & Goldenberg, A. (2019). Do no harm: a roadmap for responsible machine learning for health care. Nature Medicine, 25(9), 1337-1340.

World Health Organization. (2021). Ethics and governance of artificial intelligence for health. Geneva: World Health Organization.

CONTRIBUIÇÃO DO(A) AUTOR(A) (TAXONOMIA CREDIT)

Antonio Pires Barbosa: Conceituação (Conceptualization); Metodologia (Methodology); Investigação (Investigation); Análise formal (Formal analysis); Curadoria de dados (Data curation) – não aplicável, pois não foram utilizados dados empíricos; Redação do manuscrito original (Writing – original draft); Redação – revisão e edição (Writing – review & editing); Supervisão (Supervision); Aquisição de financiamento (Funding acquisition) – não aplicável, sem financiamento externo. Registro simulado para teste técnico da plataforma OJS.

DECLARAÇÕES EDITORIAIS

Financiamento: não recebeu financiamento externo.

Conflito de interesses: não se aplica.

Uso de inteligência artificial: não se aplica.

DISPONIBILIDADE DOS DADOS: NÃO SE APLICA; NÃO FORAM UTILIZADOS DADOS EMPÍRICOS.