Justiça Algorítmica na Saúde: Uma Revisão sobre Detecção e Avaliação de Impactos dos Vieses em Aprendizado de Máquina

Bianca Matos de Barros[1], **Julia Mombach Da Silva**[1],
João Gabriel Zandoná[1], **Mariana Recamonde-Mendoza**[1,2]

[1]Instituto de Informática – Universidade Federal do Rio Grande do Sul (UFRGS)
Caixa Postal 15.064 – 91501-970 – Porto Alegre - RS – Brasil

[2]Núcleo de Bioinformática – Hospital de Clínicas de Porto Alegre (HCPA)
Av. Protásio Alves, 211, Santa Cecília – 90035-903 – Porto Alegre - RS – Brasil

{bmbarros, jmsilva, jgzandona, mrmendoza}@inf.ufrgs.br

Abstract. *The use of machine learning in healthcare has grown, but algorithmic biases can compromise the fairness and reliability of predictions. In this review, we analyzed 56 studies published between 2020 and 2022 to investigate how biases in predictive health models were identified and measured, as well as their potential impact. Our findings indicate that fairness and interpretability metrics have been underused, and systematic approaches to ensuring equity remained insufficient. The identified biases may disproportionately harm minority patient groups by increasing diagnostic errors, reducing treatment effectiveness, and limiting access to essential support and resources.*

Resumo. *O uso de aprendizado de máquina na área da saúde tem crescido, mas vieses algorítmicos podem comprometer a equidade e a confiabilidade das predições. Nesta revisão, analisamos 56 estudos publicados entre 2020 e 2022 para investigar como os vieses em modelos preditivos para a saúde foram identificados e quantificados, e qual seu potencial impacto . Nossos resultados indicam que as métricas de equidade e interpretabilidade foram pouco exploradas e ainda faltam abordagens mais sistemáticas para garantir predições justas. Os vieses identificados podem prejudicar desproporcionalmente grupos minoritários de pacientes, aumentando erros diagnósticos, reduzindo a eficácia dos tratamentos e restrigindo o acesso a suporte e recursos essenciais.*

1. Introdução

A inteligência artificial (IA) tem um grande potencial de transformação no setor da saúde: as aplicações possíveis incluem diagnósticos de alta acurácia, tratamentos personalizados, identificação de pacientes elegíveis para estudos clínicos, e redução de custos, dentre muitos outros. Contudo, desafios como insuficiência de capacidade tecnológica, dificuldades com regulamentações e políticas, bem como com o cenário ético envolvendo o uso da IA fazem com que sua adoção nesta área seja mais lenta do que em outros setores. Fatores como necessidade de adaptação cultural e dificuldades na compreensão da tecnologia também podem contribuir para a demora na adoção de IA na área da saúde [Aldwean and Tenney 2023, Lin et al. 2024].

Um fator que pode comprometer a confiança dos usuários é a existência de vieses nos modelos computacionais. Aqui, definimos viés como uma inclinação ou preconceito

nas decisões de um sistema de IA de forma a favorecer ou desfavorecer uma pessoa ou grupo de uma maneira considerada injusta, especialmente baseando-se em características demográficas como raça, gênero, idade, dentre outras [Ntoutsi et al. 2020].

No aprendizado de máquina (AM), um subcampo da IA muito utilizado na área da saúde, os modelos realizam predições com base na representação automática de conhecimento incorporado em bancos de dados digitais [Faceli et al. 2021]. Estes bancos de dados registram a prestação de serviços em nossos sistemas de saúde, cujas disparidades são amplamente documentadas. A influência de desigualdades estruturais, diferenças no acesso aos cuidados de saúde e preconceitos dos tomadores de decisão afeta os padrões de prestação de serviço de formas que tendem a ser reproduzidas e até exacerbadas pelos modelos computacionais caso a metodologia de desenvolvimento não inclua a prevenção e a mitigação deste tipo de viés [Silva 2022].

Neste trabalho, realizamos uma revisão de literatura com objetivo de caracterizar os vieses em modelos de AM na área da saúde, identificando os principais tipos de vieses, as aplicações e algoritmos onde eles são encontrados, e as estratégias de justiça, interpretabilidade e explicabilidade que podem ser usadas para detectar sua presença. Esta caracterização pretende contribuir com o desenvolvimento de técnicas de mitigação destes vieses e promover a adoção de metodologias robustas, de forma a gerar sistemas justos e confiáveis e a reduzir as barreiras na adoção da IA na área da saúde.

Este artigo está organizado da seguinte forma: a Seção 2 apresenta trabalhos relevantes para a área de pesquisa, e diferencia esta revisão em relação à literatura existente; a Seção 3 detalha o protocolo de revisão; a Seção 4 apresenta e discute os resultados encontrados e a Seção 5 sumariza as principais contribuições e limitações deste estudo, e aponta os possíveis trabalhos futuros.

2. Trabalhos relacionados

Selecionamos alguns trabalhos relacionados à proposta desta revisão, que trazem discussões importantes para a área. Para compreensão conceitual da justiça em AM, o trabalho de [Tang et al. 2023] faz-se interessante, pois aborda as métricas de justiça e o papel da causalidade na detecção e mitigação de vieses, além de propor uma estrutura para orientar a análise de justiça algorítmica.

Adicionalmente, [Cirillo et al. 2020] avaliam as lacunas nas tecnologias de IA biomédicas atuais, discutindo a importante diferença entre vieses desejados e indesejados. Vieses desejados são aqueles que usam as diferenças entre as populações de forma a permitir diagnósticos mais precisos e tratamentos personalizados, tornando os cuidados de saúde mais eficazes. Já os vieses indesejados são aqueles que podem perpetuar desigualdades e discriminações, ampliando preconceitos existentes e levando a resultados negativos para certos grupos. Entender e saber diferenciar entre esses tipos de viés é fundamental para garantir resultados justos.

Outras revisões abordando vieses em modelos de AM em domínios diversos, sem restrição à área da saúde, também foram importantes por reunir conceitos e estatísticas sobre vieses, métricas de justiça e estratégias de mitigação encontrados na literatura, oferecendo um ponto de partida para a investigação do cenário específico das aplicações de saúde [Mehrabi et al. 2021, Siddique et al. 2023]. Da mesma forma, revisões relacionadas à identificação e mitigação de vieses em subáreas específicas da saúde, embora

mais restritas, também foram relevantes para familiarização com o domínio de aplicação [Chen and Marrero 2024, Perets et al. 2024].

Também encontramos uma revisão extensa avaliando um conjunto de 341 artigos sobre métodos de mitigação de viés. Essa revisão inclui estudos da área da saúde e aborda aspectos como conjuntos de dados e métricas de justiça, além de realizar um benchmarking. Contudo, o trabalho foca somente em métodos de mitigação, não abrangendo artigos que detectam viés sem mitigá-lo. Ele também se restringe a dados tabulares, não abrangendo estudos que usem imagem médica ou outros tipos de dados [Hort et al. 2024].

Nosso diferencial em relação à literatura existente é o objetivo de caracterizar os vieses específicos dos modelos de AM utilizados na área da saúde, abrangendo tanto problemas de classificação quanto de regressão e sem restrições quanto aos dados de entrada, abrangendo trabalhos com dados tabulares, imagem médica, informações de áudio ou vídeo, dentre outras mídias. Nosso foco está na descrição dos vieses encontrados, independente do uso de métodos de mitigação. Quando aplicável, também coletamos informações sobre as estratégias de interpretabilidade e explicabilidade usadas, tendo em vista que elas são ferramentas com forte potencial para compreensão de como os modelos aprendem e, consequentemente, quais vieses podem estar incorporados às suas predições.

3. Protocolo de Revisão

Nossa revisão foi inspirada no framework PRISMA 2020 [Page et al. 2021]. O grupo de autores é formado por pesquisadores da computação, incluindo uma docente com título de doutora, uma doutoranda e dois estudantes de graduação. Todos atuaram como revisores.

Diante de um grande número de artigos encontrados na busca inicial e de um corpo reduzido de revisores, algumas restrições foram impostas para viabilizar a execução do estudo e manter a carga de trabalho sob controle. As principais restrições foram a redução do número mínimo de revisores para um por artigo, tanto na triagem quanto na extração de informações, e a seleção do inglês como único idioma para os estudos selecionados. Estas restrições classificam este trabalho como uma Revisão Rápida [Stevens et al. 2024]

Para reduzir o potencial viés de seleção introduzido pela redução do número mínimo de revisores, utilizamos estratégias para apoiar a padronização de decisões e a rastreabilidade dos resultados [Mantsiou et al. 2023]. Estas estratégias incluíram a definição clara de critérios de inclusão e exclusão de resumos e a preparação dos revisores antes da triagem, além da escolha do Rayyan como plataforma principal, por possibilitar trabalho colaborativo, revisão blindada (na qual os revisores não tem acesso às decisões uns dos outros) e manutenção de histórico das avaliações [1] [Ouzzani et al. 2016].

3.1. Questões de Pesquisa

É comum que estudos avaliando modelos de AM se refiram a diversas tendências estatísticas como vieses, o que pode incluir conceitos como viés de seleção, viés de otimismo ou viés de *scanner* em imagem médica. Aqui, focamos nos vieses indesejados [Cirillo et al. 2020] e restringimos nosso interesse àqueles explicitamente ligados a predições injustas e ao preconceito contra grupos minoritários. A detecção destes vieses geralmente envolve atributos demográficos como raça, gênero ou idade, dentre outros. As questões de pesquisa são:

[1] https://www.rayyan.ai/

1. Como o viés ou a injustiça nas previsões feitas por modelos de AM para a saúde foram medidos, quantificados ou detectados?
2. Que tipos de viés foram discutidos no contexto de modelos de AM para a saúde?
3. Qual é o impacto do viés no poder preditivo dos modelos de AM? Como ele afeta grupos minoritários, introduzindo ou agravando disparidades na saúde?

3.2. Estratégia de busca

Consideramos trabalhos que desenvolveram ou validaram modelos de AM para prever desfechos relacionados à saúde e que identificaram, quantitativa ou qualitativamente, o problema do viés em seus modelos, avaliando seu impacto ou propondo estratégias de mitigação. Definimos palavras-chave de acordo com estas considerações e buscamos artigos publicados desde 1 de janeiro de 2020 nas bases de conhecimento Embase[2], PubMed[3] e DBLP[4], sendo esta última focada em publicações na área de ciência da computação.

Incluímos tanto artigos de periódicos quanto artigos completos em anais de conferências. Também incluímos a literatura cinza, composta por documentos produzidos por diversos setores com qualidade suficiente para serem mantidos por bibliotecas ou repositórios institucionais, mas que não são controlados por editoras comerciais [Schöpfel 2011]. Ela é uma fonte importante de evidências para revisões de literatura e pode ajudar a reduzir o efeito do viés de publicação. Buscamos literatura cinza na forma de preprints indexados na Europe PMC. Além disso, o DBLP indexa artigos na categoria de ciência da computação no ArXiv, um repositório de preprints de acesso aberto da Universidade Cornell muito usado na área. A coleta de artigos foi concluída em 26 de outubro de 2022. As strings de busca estão disponíveis em repositório público destinado a fornecer o material suplementar deste trabalho [5].

3.3. Triagem de artigos

Após a realização da estratégia de busca, os trabalhos encontrados foram carregados no Rayyan. Os artigos duplicados foram removidos, e cada participante recebeu um convite para ingresso na plataforma, estando apto a atuar na revisão após o registro individual. Os filtros do Rayyan possibilitaram identificar os artigos já avaliados ou pendentes.

Inicialmente, a elegibilidade dos trabalhos foi avaliada com base no título e no resumo, com cada artigo sendo revisado por pelo menos um avaliador no modo *blind* da plataforma, ou seja, sem que os revisores pudessem ver as decisões uns dos outros. Após a seleção inicial, uma segunda revisão foi realizada, com base na seção de Resultados, com cada artigo sendo avaliado por apenas um revisor. Nesta etapa, foi necessário desativar o modo *blind*, para que o avaliador pudesse filtrar os artigos incluídos na etapa anterior. Os artigos finalistas foram então submetidos à leitura integral para confirmação. A extração de informações foi feita com o auxílio de formulário padronizado e planilha eletrônica, complementada pelo uso dos softwares Mendeley[6] e Notion[7] para armazenamento dos textos completos e organização dos dados, respectivamente.

[2]https://www.elsevier.com/products/embase
[3]https://pubmed.ncbi.nlm.nih.gov/
[4]https://dblp.org/
[5]Repositório de material suplementar: `https://chasquebox.ufrgs.br/public/e7bdd6`
[6]`https://www.mendeley.com/`
[7]`https://www.notion.com/`

3.3.1. Critérios de inclusão

Incluímos apenas estudos que desenvolveram um modelo preditivo focado em um problema de tarefa clínica usando um algoritmo de aprendizado de máquina bem especificado, com relato claro de desempenho preditivo. Nossos critérios de inclusão são:

- O artigo especifica claramente o uso de um algoritmo de AM para o desenvolvimento do modelo;
- O artigo aborda um problema de predição clínica, incluindo, mas não se limitando a, predição do início ou estágio da doença (diagnóstico), evolução da doença (prognóstico) ou resposta ao tratamento (predição);
- O artigo detecta, discute, avalia ou aborda o problema de viés de predição em modelos baseados em AM.

3.3.2. Critérios de exclusão

Foram excluídos artigos anteriores a 2020 e que não pudessem ser recuperados pela triagem de resumo ou texto completo. Também foram excluídos trabalhos que não fossem artigos originais de pesquisa, como resumos, correspondências, revisões ou meta-análises. Adicionalmente, foram excluídos artigos que não abordassem tarefas de predição clínica a nível de paciente, não usassem métodos de AM ou não fornecessem resultados claros do desempenho preditivo obtido, bem como artigos que não detectassem, discutissem ou mitigassem de forma clara a presença de vieses indesejados.

Por fim, usamos a avaliação automática de artigos disponível no Rayyan, excluindo artigos avaliados com valor inferior a 4.5 estrelas. Esta funcionalidade de predição de resumos do Rayyan aprende com as decisões prévias dos revisores. Conforme os usuários rotulam estudos como incluídos ou excluídos, o Rayyan treina um algoritmo do tipo *Support Vector Machine* (SVM) com esses exemplos, gerando um modelo que prediz a classificação de estudos pendentes atribuindo-lhes uma pontuação de uma a cinco estrelas. O sistema atualiza iterativamente o modelo com base em novas decisões, aprimorando suas previsões até que não haja mais estudos a rotular ou até que o modelo não possa mais ser melhorado. Utilizamos a avaliação automática após a triagem manual de um conjunto inicial de artigos suficiente para treinamento do algoritmo preditor.

3.4. Análise e apresentação de resultados

A análise avaliou o perfil dos trabalhos, visando identificar tendências gerais e padrões relacionados às aplicações e metodologias, com auxílio de gráficos e tabelas. Para responder às perguntas de pesquisa, foi necessário focar nos modelos de AM, nas métricas de desempenho, nas estratégias de justiça e explicabilidade e nos vieses detectados.

4. Resultados e Discussão

A aplicação da estratégia de busca gerou um conjunto de 4760 artigos, dos quais 390 foram selecionados na primeira fase da triagem, a partir da leitura do título e do resumo. Destes 390, 56 foram selecionados após a leitura da seção de resultados, seguida pela leitura completa. Dentre os excluídos nesta etapa, a maior parte não abordava vieses relacionados a predições injustas, conforme mostrado na Tabela 1. Mais de um motivo

de exclusão podia ser atribuído a cada trabalho, de forma que a soma das quantidades de artigos por motivo de exclusão é superior ao total de artigos excluídos. A lista completa de artigos está disponível no repositório de material suplementar, junto com gráficos e informações adicionais sobre o estudos selecionados.

Tabela 1. Artigos excluídos na última etapa da revisão, por critério de exclusão.

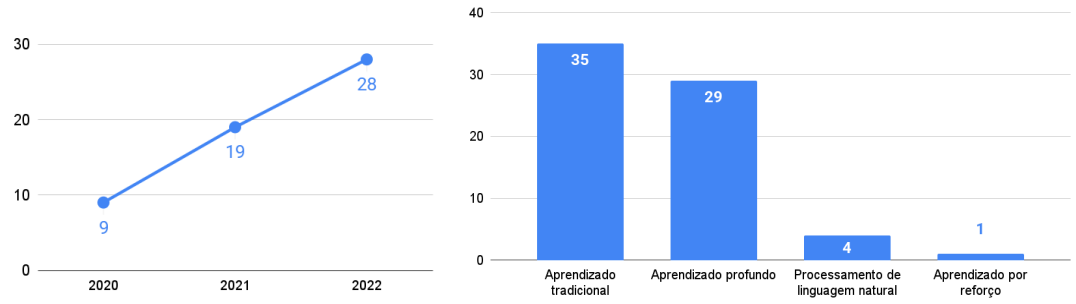
Critério de exclusão	**Quantidade**
Não há detecção, discussão ou mitigação de viés.	221
Escopo incorreto (não é modelo clínico).	51
Metodologia incorreta (não utiliza modelos de AM).	28
Fora do período de publicação previsto (anterior a 2020).	24
Tipo de publicação errado.	14
Texto completo não disponível.	7
Relatórios de desempenho preditivo insuficientes.	2
População errada (não conduzido com humanos).	1

Dentre os artigos excluídos por não detectar, discutir ou mitigar vieses, os trabalhos frequentemente abordavam tendências não relacionadas de forma direta com a predição injusta. Alguns exemplos incluem viés de seleção em estudos clínicos, viés de otimismo ou vieses relacionados a questões específicas de imagem médica. Também houve muitos casos de estudos que usaram AM para quantificar as disparidades de saúde, ou a prevalência de questões de raça/etnia, gênero ou idade como preditores, fatores de risco ou comorbidades de determinadas patologias, mas sem abordar o viés como uma diferença de desempenho do modelo entre diferentes populações.

Os outros dois motivos mais frequentes de exclusão foram os trabalhos que não se enquadravam nos critérios de escopo e metodologia. No caso do escopo, os trabalhos não apresentavam tarefas preditivas a nível de paciente, sendo os trabalhos epidemiológicos os mais comuns [Lu et al. 2021]. Com relação à metodologia incorreta, os trabalhos não usavam AM, sendo comum o uso de modelos de estatística clássica, como a regressão linear ou a regressão de Cox para estimativa de sobrevivência [Neumann et al. 2022].

Os 56 artigos selecionados formam um grupo bastante diverso e heterogêneo de aplicações e metodologias. O número de artigos publicado por ano apresentou uma tendência de crescimento expressivo entre 2020 e 2022 (Figura 1(a)), evidenciando a consolidação da área de pesquisa e o interesse crescente da comunidade científica. A maior parte dos trabalhos abordava somente um problema clínico, mas alguns estudos abordavam mais de um simultaneamente, totalizando 31 artigos (53%) realizando diagnóstico, 23 (40%) fazendo prognóstico e 4 (7%) abordando análises preditivas principalmente relacionadas ao tratamento de doenças e seus efeitos.

Para responder às perguntas de pesquisa, começamos com o tema da primeira questão: como o viés ou a injustiça nas previsões feitas por modelos de AM para a saúde foram medidos, quantificados ou detectados? A detecção dos vieses pode variar de acordo com o tipo de método e métrica de avaliação utilizado. Como pode ser visto na Figura 1(b), a maior parte dos estudos utilizou métodos de AM tradicional, com uma ampla variedade de objetivos e tipos de dados utilizados. Vários trabalhos utilizaram métodos de múltiplas categorias de aprendizado simultaneamente, de forma que a soma das quantida-



(a) Publicações por ano (b) Publicações por categoria de aprendizado

Figura 1. Número de artigos por ano de publicação e categoria de aprendizado

des por categoria supera o total de artigos selecionados. Os algoritmos de AM tradicional mais usados foram Logistic Regression (18 trabalhos) e Random Forest (13).

Dentre os trabalhos que utilizaram aprendizado profundo, houve uma predominância de estudos focados em imagem médica, com uma grande parte dos trabalhos usando dados não textuais, como radiografias, eletrocardiogramas, dados de ressonância magnética e vídeos, de forma isolada ou junto a dados textuais ou tabulares contendo informação clínica ou demográfica. Alguns estudos uniram métodos de aprendizado tradicional e profundo para objetivos diversos, em geral usando dados tabulares.

Os métodos de processamento de linguagem natural (PLN) foram usados de forma isolada ou junto a algoritmos de classificação tradicionais. Além disso, 4 estudos utilizaram o conceito de aprendizado ensemble. Por fim, o único estudo usando aprendizado por reforço tinha como objetivo propor um framework de aprendizado profundo no qual uma função especializada de recompensa foi introduzida para mitigar vieses durante o treinamento do modelo.

É importante salientar que a seleção de uma pequena quantidade de trabalhos abordando PLN, aprendizado ensemble ou aprendizado por reforço não significa que não existam trabalhos discutindo vieses com estas técnicas. Nossa hipótese é que não recuperamos muitos destes trabalhos por não haver menção explícita a eles nos termos de busca utilizados. Isso representa uma limitação de escopo em nosso estudo.

A métrica predominante para avaliação de desempenho foi a Área Sob a Curva Característica de Operação do Receptor (AUROC, *Area Under Receiver Operating Characteristic Curve*), utilizada em 43 dos 56 estudos. Precisão (14 trabalhos), Revocação ou Recall (16), F1-score (10) e Área sob a Curva Precisão-Revocação (AUPRC, *Area Under Precision-Recall Curve*, 12) também foram utilizadas. Diversas outras métricas apareceram em menor escala; algumas não foram usadas para avaliação de performance geral, mas foram relativamente frequentes na comparação de desempenho entre subgrupos, funcionando mais como métricas de justiça. Alguns exemplos são a acurácia balanceada e as taxas de verdadeiros positivos, verdadeiros negativos, falsos positivos, falsos negativos, omissão falsa e descoberta falsa, aparecendo de 2 a 10 vezes cada.

Poucas métricas específicas de justiça foram mencionadas. Dos 56 trabalhos selecionados, 39 utilizaram somente a comparação entre grupos, através das métricas de desempenho, para detectar a presença de vieses. Dentre os que usaram métricas de

justiça, 8 mencionaram a métrica de Igualdade de Oportunidades (*Equal Opportunity*) e 5 mencionaram Igualdade de Probabilidades (*Equalized Odds*). Paridade de classificação e Diferença de Paridade Estatística apareceram duas vezes cada, e Disparate Impact, Paridade Demográfica, Diferença Média de Probabilidades e Igualdade de Tratamento apareceram apenas uma vez cada. O cenário encontrado é de pouca difusão das métricas de justiça nos trabalhos que buscam detectar e discutir vieses algorítmicos.

Um dos estudos selecionados explorou a relação de diversas técnicas de interpretabilidade e explicabilidade com os conceitos e métricas de justiça algorítmica, testando um grande grupo de técnicas e demonstrando que a importância dos atributos reportada por elas pode ser usada para quantificar desigualdades nas variáveis preditoras de mortalidade, bem como contribuir para diferenças no desempenho do modelo entre grupos minoritários, de forma que é recomendável considerar interpretabilidade e justiça algorítmica de forma integrada [Meng et al. 2022].

Dentre os demais artigos selecionados nesta revisão, poucos usaram técnicas de interpretabilidade e explicabilidade. Apenas 4 e 2 trabalhos usaram as técnicas bem estabelecidas SHapley Additive exPlanations (SHAP) e Local Interpretable Model-Agnostic Explanations (LIME), respectivamente. Dentre os trabalhos com imagem médica, 2 utilizaram mapas de calor Grad-CAM e 1 utilizou mapas de saliência. Sete estudos avaliaram as importâncias de atributos usando funções nativas dos classificadores, e apenas 2 propuseram técnicas relacionadas à explicabilidade, sendo uma relacionada ao tratamento de variáveis de confusão e outra relacionada a avaliação de importância de atributos em eletrocardiogramas. O número baixo de artigos utilizando estas técnicas também pode ter relação com a ausência destas técnicas em nossos termos de busca, o que também configura uma restrição de escopo.

A integração entre interpretabilidade, explicabilidade e justiça ainda precisa ser bastante explorada, pois é uma ferramenta valiosa para compreensão do aprendizado dos modelos e discernimento entre correlações de importância clínica e vieses induzidos pelos dados ou modelos. Ao tornar visível a lógica por trás das decisões dos modelos, a interpretabilidade e a explicabilidade viabilizam ajustes mais precisos nos dados e na modelagem, enquanto as métricas de justiça ajudam a monitorar o impacto dessas intervenções sobre diferentes grupos populacionais. Juntas, essas abordagens contribuem para o desenvolvimento de sistemas mais transparentes, justos e clinicamente confiáveis.

Os vieses em geral são aprendidos pelo modelo através das diferenças entre padrões presentes nas variáveis de entrada. Estudos utilizando dados tabulares, muito comuns em todas as categorias de métodos de AM, podem conter representações de diversas desigualdades. Determinados grupos raciais ou de gênero, por exemplo, podem ter menos acesso a hábitos saudáveis, consultas médicas, procedimentos e outros recursos de saúde, bem como estar mais frequentemente associados a um determinado seguro de saúde, correlações estas que são exploradas de forma intrínseca pelos modelos de AM.

Nos estudos utilizando imagem médica, essas correlações também podem ser exploradas pelos modelos, que podem associar marcadores de determinados grupos, como a densidade óssea reduzida em pessoas mais velhas ou as diferenças na estatura média entre homens e mulheres, a determinados desfechos. Nos estudos utilizando PLN, os modelos podem aprender as representações sociais reproduzidas pelos profissionais de saúde, que

podem ter condutas diferenciadas entre grupos, fazendo diferentes perguntas, suposições e até mesmo usando termos associados a determinadas populações. Além de representar um risco a populações minoritárias, esse tipo de aprendizado pelos modelos de AM também é feito a partir de associações sem relevância clínica, prejudicando a geração de conhecimento por meio de achados espúrios.

Seguimos avaliando o tema da segunda pergunta de pesquisa: Que tipos de viés foram discutidos no contexto de modelos de aprendizado de máquina para a saúde? Como exibido na Figura 2, os vieses raciais e de gênero foram os mais frequentes nos estudos selecionados. Os trabalhos abordando este tipo de disparidade em geral utilizam definições padronizadas a partir da autodeclaração de raça, etnia ou sexo biológico, e frequentemente encontram diferenças de desempenho entre grupos brancos e não-brancos, e entre homens e mulheres. Os vieses relacionados a faixa etária também foram bastante presentes, em geral afetando grupos menos representados nas bases de dados, como idosos ou crianças.

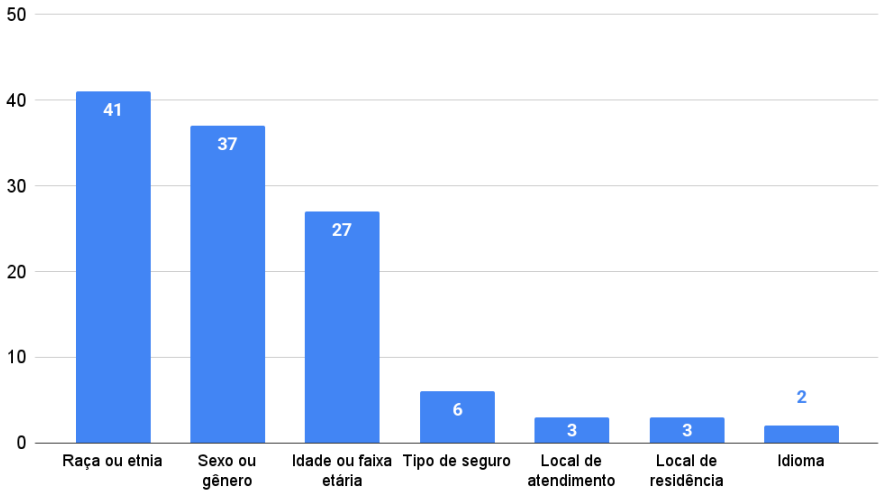


Figura 2. Número de artigos por tipo de viés indesejado avaliado.

Em um número menor de casos, também foram detectados vieses relacionados ao tipo de seguro, local de atendimento ou residência ou, ainda, idioma falado pelo paciente. Todas estas variáveis podem atuar como *proxies* para grupos desfavorecidos em determinados contextos, em especial o tipo de seguro, que é considerado em vários estudos como um indicador de status socioeconômico do paciente.

A última pergunta de pesquisa se refere a efeitos práticos: qual é o impacto do viés no poder preditivo dos modelos de AM? Como ele afeta grupos minoritários, introduzindo ou agravando disparidades na saúde?

Para avaliar a questão de forma mais consistente, os artigos foram agrupados em duas grandes áreas. Uma área foi a assistência à saúde, com cerca de 34% dos trabalhos, trazendo tarefas preditivas relacionadas a assistência hospitalar. O prognóstico de mortalidade foi a mais frequente, junto a outros prognósticos, como admissão, tempo de estadia, e necessidade de unidade de terapia intensiva ou ventilação mecânica. A outra área, com os 66% restantes, foi composta por trabalhos relacionados a condições de saúde, com tarefas preditivas em patologias diversas. Patologias respiratórias, principalmente Covid-19, foram os casos mais frequentes (Figura 3). Em oito casos, os trabalhos abordavam

ambas as áreas simultaneamente, e foram associados aos dois grupos.

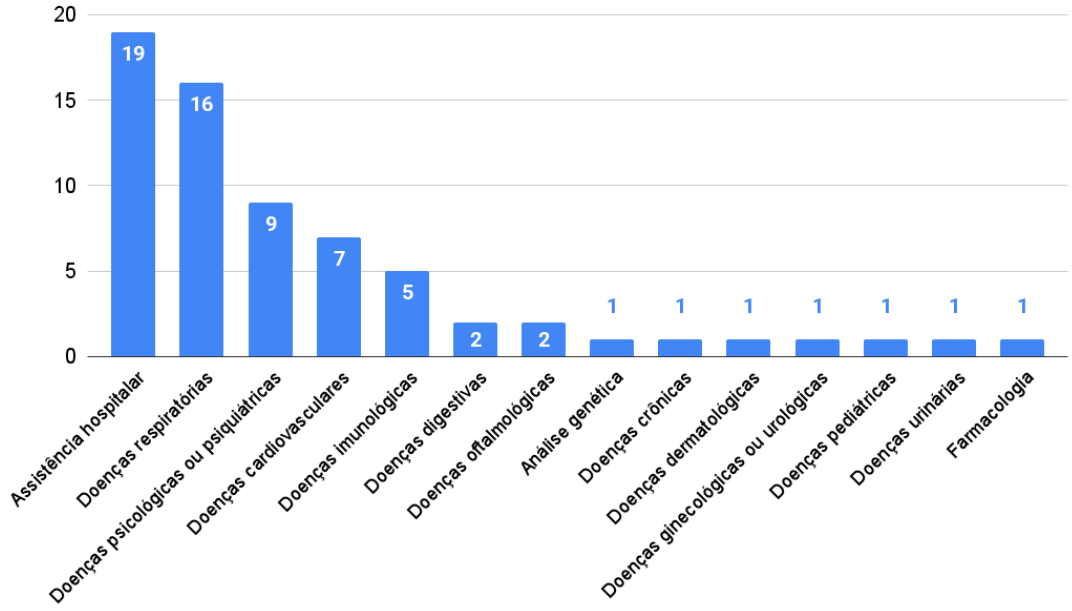


Figura 3. Número de artigos por área

Nos modelos aplicados a condições de saúde, as tarefas preditivas predominantes estão relacionadas ao diagnóstico e à previsão de evolução ou resposta ao tratamento. Desempenho reduzido para um determinado grupo neste cenário pode significar maiores taxas de erro de diagnóstico, de estimativas de risco equivocadas e menor eficácia na prescrição de tratamentos.

Nos modelos aplicados à assistência hospitalar, as tarefas preditivas estão geralmente associadas a estimativas de risco, de mortalidade ou outros prognósticos importantes. Desempenho reduzido para um determinado grupo neste cenário pode significar atraso ou falta de suporte adequado e retenção de recursos em casos potencialmente críticos, como na necessidade de ventilação mecânica ou de admissão na UTI.

Em ambos os casos, o viés algorítmico alimenta ciclos onde o atendimento desigual é reproduzido, e ao mesmo tempo, gera novos registros que podem ser usados pra treinamento de sistemas de IA que irão aprender novamente os padrões discriminatórios, perpetuando e eventualmente agravando as disparidades na saúde. É indispensável adicionar etapas de auditoria e mitigação de vieses indesejados para a promoção da equidade na saúde, o desenvolvimento de modelos de AM confiáveis, e consequentemente, a redução da barreira de entrada da IA na área da saúde.

Nosso trabalho contribui para uma caracterização clara dos vieses em modelos de AM na área da saúde, bem como das métricas de justiça, interpretabilidade e explicabilidade que podem ser usadas para detectar sua presença. Ele apresenta duas limitações principais. A primeira refere-se ao fato de cada artigo ter sido avaliado por apenas um revisor, o que pode introduzir viés de seleção na triagem e categorização dos estudos. Para mitigar esse potencial viés, investimos em padronização e rastreabilidade rigorosas dos resultados. A segunda limitação está relacionada ao escopo das estratégias de busca, que, por não incluírem termos específicos associados a técnicas de explicabilidade e pa-

radigmas específicos de AM, como ensemble, PLN e aprendizado por reforço, podem ter sub-representado estudos relevantes nessas áreas. Ainda assim, os termos utilizados foram definidos com base em descritores amplamente reconhecidos, visando captar uma amostra representativa dos estudos relevantes na área.

5. Conclusão

Nossa revisão revelou um crescente interesse científico nos vieses algorítmicos em modelos de AM para a saúde. No entanto, identificamos limitações na detecção e quantificação desses vieses, com a maioria dos estudos focando em métricas padrão de desempenho, enquanto métricas específicas de justiça algorítmica e interpretabilidade foram pouco exploradas.

Para pesquisas futuras, pretendemos atualizar o conjunto de estudos analisados para incluir a literatura mais recente, e adicionar informações sobre bases de dados, pacotes de software mais utilizados e técnicas de mitigação. Também ressaltamos a necessidade de continuar avaliando estudos com diferentes tipos de dados, além dos tabulares, para uma compreensão mais abrangente do problema. Por fim, salientamos que, além de mitigar vieses, os modelos de AM podem servir como ferramentas poderosas para auditar desigualdades nos sistemas de saúde e fundamentar políticas públicas voltadas à promoção da equidade.

Agradecimentos

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001 e da Fundação de Amparo à Pesquisa do Estado do Rio Grande do Sul (FAPERGS) [21/2551-0002052-0 (Projeto MARCS e 22/2551-0000390-7 (Projeto CIARS)]. M. Recamonde-Mendoza é parcialmente financiada pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) [308075/2021-8].

Referências

Aldwean, A. and Tenney, D. (2023). Artificial intelligence in healthcare sector: a literature review of the adoption challenges. *Open Journal of Business and Management*, 12(1):129–147.

Chen, Z. and Marrero, W. J. (2024). A survey on optimization and machine-learning-based fair decision making in healthcare. *medRxiv*, pages 2024–03.

Cirillo, D., Catuara-Solarz, S., Morey, C., Guney, E., Subirats, L., Mellino, S., Gigante, A., Valencia, A., Rementeria, M. J., Chadha, A. S., et al. (2020). Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare. *NPJ Digital Medicine*, 3(1):1–11.

Faceli, K., Lorena, A. C., Gama, J., Almeida, T. A. d., and Carvalho, A. C. P. d. L. F. d. (2021). *Inteligência artificial: uma abordagem de aprendizado de máquina*. LTC.

Hort, M., Chen, Z., Zhang, J. M., Harman, M., and Sarro, F. (2024). Bias mitigation for machine learning classifiers: A comprehensive survey. *ACM Journal on Responsible Computing*, 1(2):1–52.

Lin, N., Paul, R., Guerra, S., Liu, Y., Doulgeris, J., Shi, M., Lin, M., Engeberg, E. D., Hashemi, J., and Vrionis, F. D. (2024). The frontiers of smart healthcare systems. In *Healthcare*, volume 12, page 2330. MDPI.

Lu, L., Anderson, B., Ha, R., D'Agostino, A., Rudman, S. L., Ouyang, D., and Ho, D. E. (2021). A language-matching model to improve equity and efficiency of covid-19 contact tracing. *Proceedings of the National Academy of Sciences*, 118(43):e2109443118.

Mantsiou, C., Liakos, A., Mainou, M., Papanas, N., Tsapas, A., and Bekiari, E. (2023). A simple guide to systematic reviews and meta-analyses. *The International Journal of Lower Extremity Wounds*, page 15347346231169842.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54(6):1–35.

Meng, C., Trinh, L., Xu, N., Enouen, J., and Liu, Y. (2022). Interpretability and fairness evaluation of deep learning models on mimic-iv dataset. *Scientific Reports*, 12(1):7166.

Neumann, J. T., Thao, L. T., Murray, A. M., Callander, E., Carr, P. R., Nelson, M. R., Wolfe, R., Woods, R. L., Reid, C. M., Shah, R. C., et al. (2022). Prediction of disability-free survival in healthy older people. *Geroscience*, 44(3):1641–1655.

Ntoutsi, E., Fafalios, P., Gadiraju, U., Iosifidis, V., Nejdl, W., Vidal, M.-E., Ruggieri, S., Turini, F., Papadopoulos, S., Krasanakis, E., et al. (2020). Bias in data-driven artificial intelligence systems—an introductory survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(3):e1356.

Ouzzani, M., Hammady, H., Fedorowicz, Z., and Elmagarmid, A. (2016). Rayyan—a web and mobile app for systematic reviews. *Systematic Reviews*, 5:1–10.

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., et al. (2021). The prisma 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, 372.

Perets, O., Stagno, E., Yehuda, E. B., McNichol, M., Celi, L. A., Rappoport, N., and Dorotic, M. (2024). Inherent bias in electronic health records: A scoping review of sources of bias. *medRxiv*.

Schöpfel, J. (2011). Towards a prague definition of grey literature. *The grey journal*, 7(1):5–18.

Siddique, S., Haque, M. A., George, R., Gupta, K. D., Gupta, D., and Faruk, M. J. H. (2023). Survey on machine learning biases and mitigation techniques. *Digital*, 4(1):1–68.

Silva, T. (2022). *Racismo algorítmico: inteligência artificial e discriminação nas redes digitais*. Edições Sesc SP.

Stevens, A., Hersi, M., Garritty, C., Hartling, L., Shea, B. J., Stewart, L. A., Welch, V. A., and Tricco, A. C. (2024). Rapid review method series: interim guidance for the reporting of rapid reviews. *BMJ Evidence-based Medicine*.

Tang, Z., Zhang, J., and Zhang, K. (2023). What-is and how-to for fairness in machine learning: A survey, reflection, and perspective. *ACM Computing Surveys*, 55(13s):1–37.