Robô Laura
Auditoria algorítmica

Estudo sobre o sistema Laura de previsão do risco de deterioração clínica

eticas BID

Robô Laura
Auditoria algorítmica

Estudo sobre o sistema Laura de previsão do risco de deterioração clínica

Dezembro de 2021

https://www.iadb.org/

Tradução de Mariana Fagundez, otraspalabras.com.

Índice

Laura é um **Robô Cognitivo/ Gerenciador de Riscos** que atua na **identificação precoce dos riscos de deterioração clínica.**

INTRODUÇÃO

Este documento apresenta o Relatório Final da auditoria algorítmica do sistema Laura, realizada pela *Eticas Research and Consulting*.[1] Esta auditoria não aborda apenas a justiça algorítmica no modelo de processamento automatizado; também inclui uma avaliação da desejabilidade, aceitabilidade e gerenciamento dos dados nesse sistema. No documento, os resultados dessas análises são desenvolvidos com base na descrição de seu modelo e na análise de seu âmbito social de implementação.

O sistema Laura é um **Robô Cognitivo/Gerenciador de Riscos** que atua na **identificação precoce dos riscos de deterioração clínica.** Em atividade desde 2016, o sistema Laura já analisou mais de 8,6 milhões de consultas em 40 centros clínicos e hospitalares de diversos estados do Brasil. O sistema vem mudando seu modelo, passando de uma abordagem focada na identificação do risco de apresentar sepse em pacientes internados para uma mais integral, na qual a avaliação é baseada no risco de deterioração clínica e morte, com base em parâmetros semelhantes. Esta **auditoria examina o aplicativo Laura, em sua versão 1.0**, criado em 2017.

O principal objetivo do sistema Laura é alertar precocemente a deterioração clínica suscetível à morte, com o efeito de reduzir a mortalidade e os custos dos serviços hospitalares por meio de análise preditiva[2]. É um sistema de Inteligência Artificial que oferece uma classificação do risco de piora clínica do paciente, após uma análise dos indicadores das últimas cinco coletas de sinais vitais do paciente. Esse sistema de previsão de risco foi contrastado com o **sistema de pontuação *Modified Early Warning Score*[3] (MEWS),** usado como padrão para detecção precoce de deterioração clínica (Kobylarz et al., 2020). A plataforma inteligente está atualmente conectada na nuvem a mais de 40 hospitais brasileiros que possuem diferentes Prontuários Eletrônicos dos Pacientes (PEPs).

A **auditoria algorítmica do sistema Laura** concentrou-se em **explorar possíveis riscos de viés ou discriminação algorítmica** nos resultados oferecidos pelo sistema e efetivamente traduzidos em intervenções clínicas por parte das equipes hospitalares. Nesse sentido, vale ressaltar que, entre os produtos oferecidos pela Laura, que incluem ferramentas de Detecção de Deterioração Clínica, Atenção Primária, Gestão de Protocolos e Perfil Epidemiológico, centraremos a nossa atenção nas ferramentas de detecção de deterioração clínica e seus gestores de informação.

Para isso, nas **Seções 2 "Estado da arte" e 3 "Contexto social"** deste documento, a nossa análise enquadra o sistema Laura em um estado da arte referente a sistemas automatizados de previsão de deterioração clínica e situa-o em seu contexto socioeconômico e cultural. Essa análise é realizada por meio de um estudo da literatura, estatísticas relevantes e entrevistas com desenvolvedores do sistema, bem como com equipes clínicas e de enfermagem responsáveis por sua implementação em hospitais[4]. Com base nisso, estabelecem-se hipóteses de viés algorítmico para medição quantitativa em outra fase da auditoria, por meio dos resultados oferecidos pelo sistema em um período específico.

1 A equipe de pesquisa foi composta por Emma López e Mariano Martín Zamorano. A equipe contou com a assessoria do Dr. Carlos Castillo, da Universidade Pompeo Fabra.

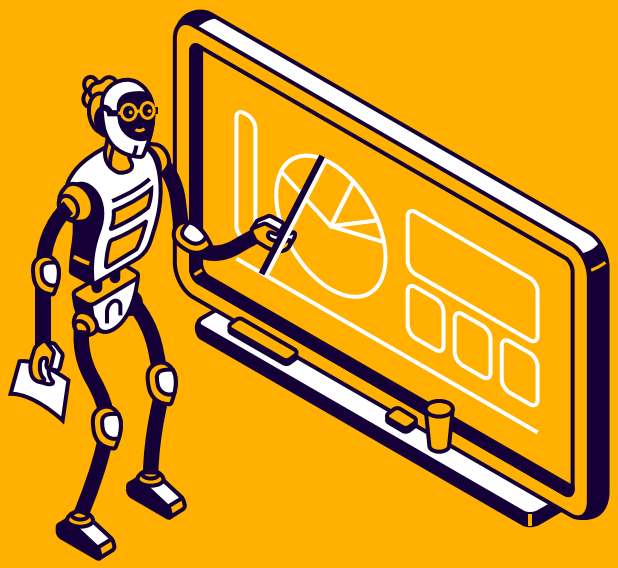
2 Veja a apresentação em https://www.laura-br.com/en/

3 As pontuações de alerta precoce (MEWS, na sigla em inglês) foram desenvolvidas para melhorar os mecanismos de detecção de deterioração com base nos parâmetros fisiológicos dos pacientes nos quartos de um hospital (Morgan et al., 1997). O uso de dados sobre deterioração intra-hospitalar e parada cardíaca demonstrou ser precedido por um período de aumento das anormalidades dos sinais vitais (Williams, 2017). Mais informações sobre o *EWS em https://en.wikipedia.org/wiki/Early_warning_score

4 Essas primeiras entrevistas incluíram: entrevista 1, com a equipe técnica (17/03/2021), e entrevista 2, com a diretoria e a coordenação assistencial do sistema (28/03/2021).

Na **Seção 4 “Estratégia Metodológica”**, o documento descreve a **metodologia estabelecida** para medir discriminação e justiça algorítmica no sistema. Essa metodologia foi proposta como uma exploração focada na coleta de evidências indiretas sólidas sobre viés algorítmico para as variáveis sexo biológico e idade. Essa abordagem não esgotará todas as formas de avaliação desse fenômeno, mas fornecerá instrumentos para consideração e monitoramento mais abrangente no futuro. A referida metodologia destina-se, ainda, ao estudo de seu impacto social em termos de usabilidade, desejabilidade e aceitabilidade, bem como à análise do tratamento de dados pessoais.

Por fim, o relatório apresenta os **resultados da análise qualitativa e quantitativa** do sistema e as recomendações relacionadas a ele. Esses resultados serão apresentados nas **Seções 5 “Resultados do estudo operacional e de aceitabilidade”, 6 “Resultados da análise algorítmica” e 7 “Conclusões e recomendações”**, que abordam as quatro principais dimensões da auditoria, a proteção de dados pessoais, a aceitabilidade e usabilidade da Laura, e a justiça algorítmica na atribuição de risco de deterioração clínica, fechando com algumas conclusões que englobam recomendações derivadas específicas.



1. COMO O SISTEMA FUNCIONA

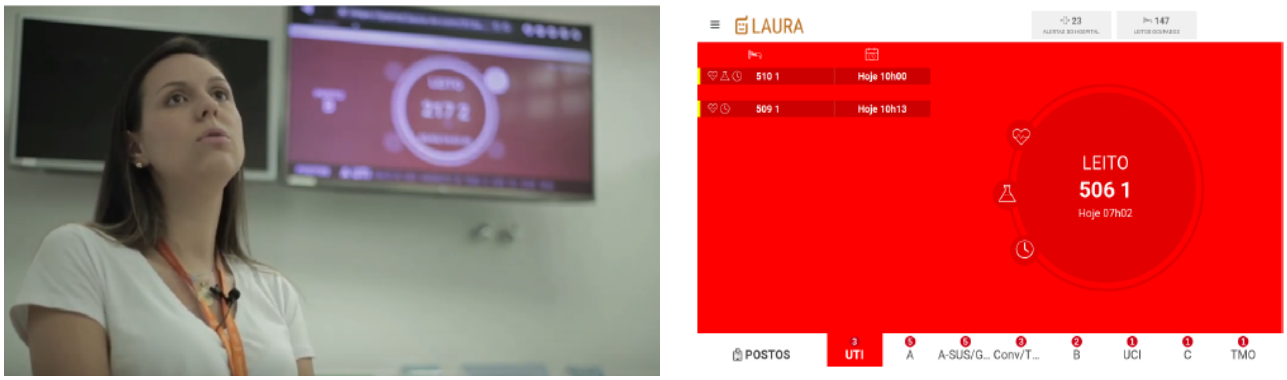
1. COMO O SISTEMA FUNCIONA

PREVISÃO DE DETERIORAÇÃO CLÍNICA

O Robô Laura é um sistema especializado de **avaliação da deterioração clínica.** Consiste em uma plataforma inteligente conectada na nuvem a mais de 40 hospitais brasileiros. Esses hospitais possuem diferentes Prontuários Eletrônicos dos Pacientes (PEPs) (Kobylarz et al., 2020) que cada hospital armazena em seus próprios bancos de dados (entrevista virtual, 17/3/2021). Esses PEPs não são padronizados no âmbito inter-hospitalar.

Usando Inteligência Artificial (IA) e aprendizado de máquina, o sistema fornece alertas antecipados à equipe de atenção médica na forma de taxa de risco e outras informações sobre a condição do paciente. Essas informações, que refletem a condição clínica do paciente em tempo real (Kobylarz et al., 2020), são monitoradas por meio de um painel ao vivo, onde os alertas médicos são exibidos à medida que ocorrem (Figura 1). Por meio dessa comunicação (que também inclui diretrizes sobre como a equipe assistencial deve agir ao identificar as alterações), o robô indica quais pacientes podem estar em **alto, médio e baixo riscos de deterioração clínica.**

Figura 1. Painel do sistema Laura



Fonte: Laura.

O funcionamento do sistema Laura, passo a passo, é resumido por Kalil et al. (2018: 311) do seguinte modo:

I. **Realiza a mineração de dados**, remotamente, em todos os bancos de dados e equipamentos de geração de dados do hospital.

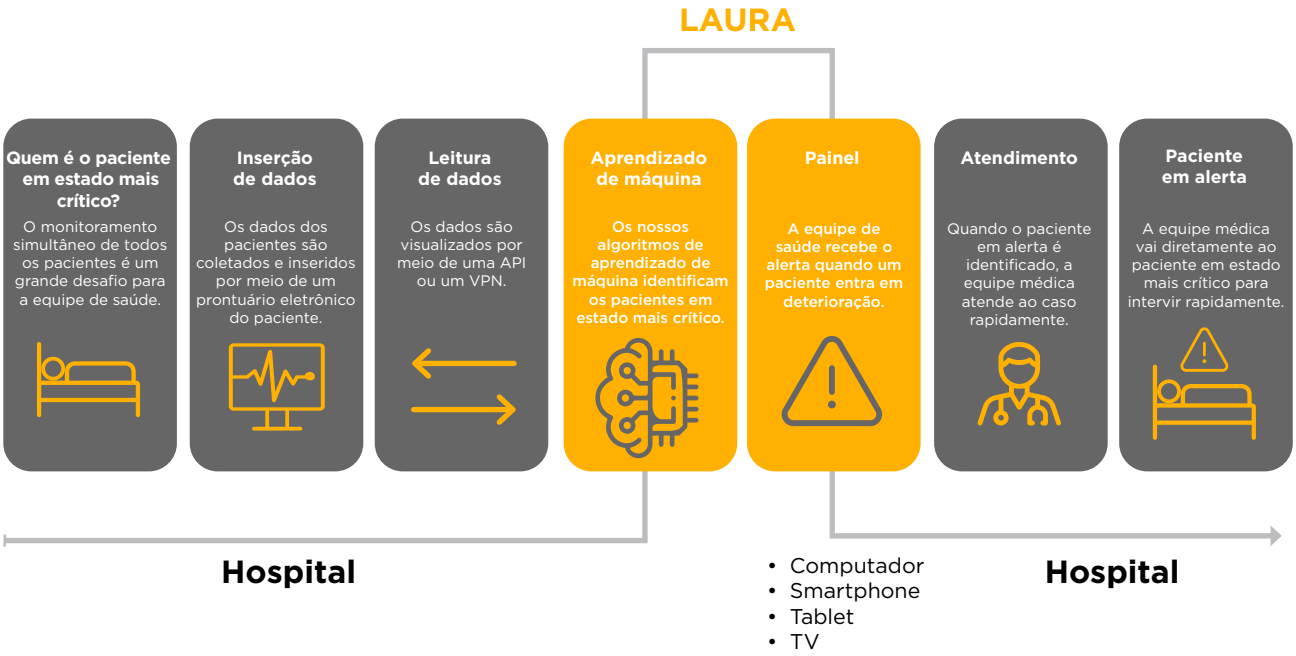
II. **Classifica** registros anômalos, inconsistentes e defeituosos.

III. **Avalia** esses dados para gerar alarmes de risco para cada paciente, com a intervenção do médico especialista em algoritmos.

IV. **Organiza** esses alarmes de acordo com sua frequência e importância em áreas de risco. Isso é traduzido visualmente para a equipe assistencial por painéis de gerenciamento de visualização instalados na enfermaria do hospital.

V. **Ativa** autonomamente a comunicação funcional do espectro quando a área de risco mais crítica está ativa; os dados continuam a alertar sobre danos. Esta função também gerencia o envio de mensagens por SMS (Short Message Service) e e-mails aos profissionais de saúde responsáveis. Dessa forma, o sistema Laura chama a atenção para o risco captado pelo robô e antecipa a atenção que deve ser direcionada aos pacientes envolvidos.

Figura 2. Ciclo de funcionamento da Laura



Fonte: Laura.

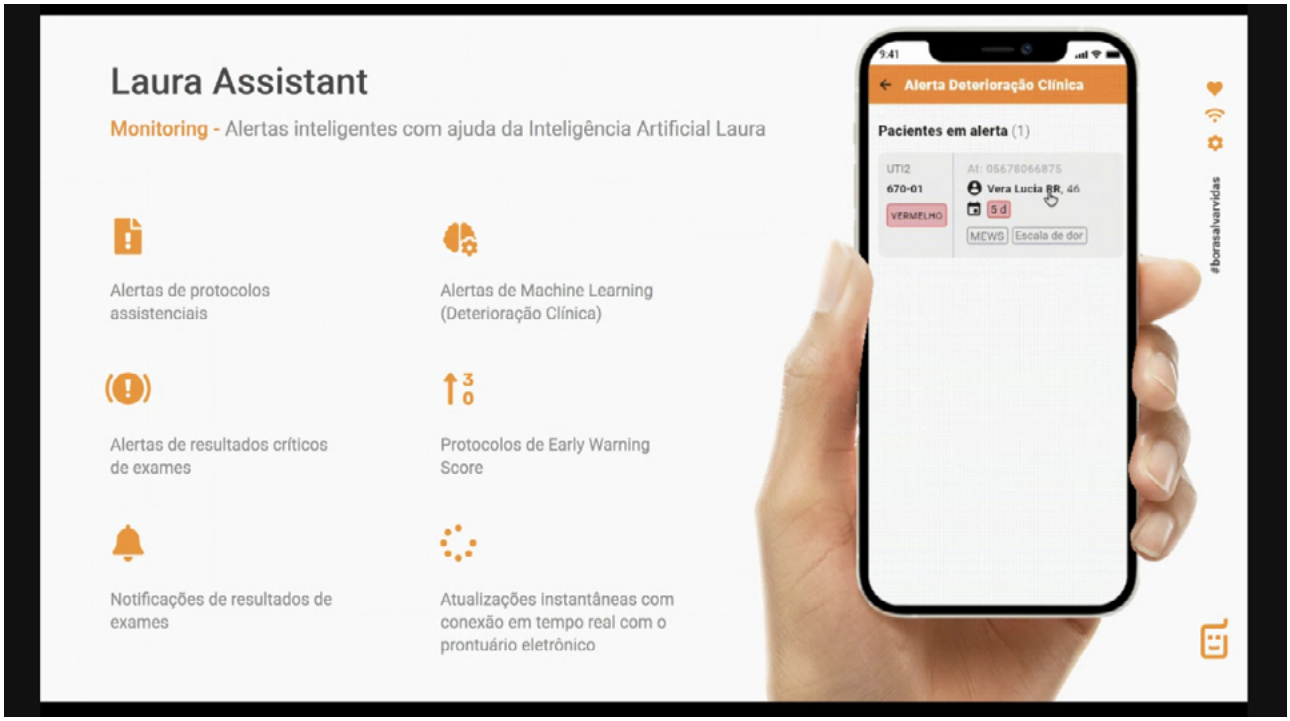
LAURA ASSISTANT

A Laura possui um **aplicativo móvel** que atualmente é seu foco de desenvolvimento, pois permite acesso mais frequente e rápido aos dados por parte da equipe médica e de enfermagem.

Por meio do aplicativo, no futuro, a equipe de saúde contará com um **relatório contínuo das pessoas que estão em risco.** O sistema desenvolvido relata diversos dados, como o número de pessoas afetadas, sua evolução ao longo do tempo e os tipos de intervenção necessários de forma mais dinâmica. Também será possível relatar a eficiência da intervenção em relação aos pacientes de maior risco (vermelho) e às taxas de atendimento dos pacientes por período. Além disso, por meio do dispositivo móvel e em comunicação com a equipe, **será possível realizar intervenções** e tomar decisões sobre o paciente em questão (a própria interface oferece opções de intervenção à equipe médica). Da mesma forma, o aplicativo oferecerá um registro contínuo de seu acompanhamento e centralizará todas as mensagens para a equipe.

A disponibilização de dispositivos móveis deve promover um acesso mais dinâmico aos dados dos pacientes e uma interação maior da equipe médica. Espera-se que essas informações gerem mais consciência sobre o desempenho das equipes e estimulem o atingimento de novas metas por meio da geração de dados estatísticos estruturados.

Figura 3. Apresentação do Laura Assistant



Fuente: Laura.

Conforme indicado em uma das nossas entrevistas (18/3/2021), a equipe do sistema Laura considera que os aspectos procedimentais são cruciais para sua efetiva utilização. Nesse sentido, o sistema está em processo de reformulação de sua metodologia e escopo de atuação, que tem passado do foco inicial na sepse para a atenção atual ao conjunto da deterioração clínica, a fim de servir como ferramenta de apoio à decisão clínica. Dessa forma, a empresa busca garantir um acompanhamento contínuo e preciso da condição do paciente, que vai além dos pontos de deterioração clínica. Nesse sentido, pretende-se oferecer apoio e capacidade administrativa ao pessoal médico para o acompanhamento de outros processos de gestão e controle hospitalar.

CARACTERÍSTICAS E BENEFÍCIOS

Atualmente, o sistema Laura tem como objetivo apoiar as atividades e o desempenho da equipe hospitalar, bem como melhorar a infraestrutura dos hospitais. Em particular, a equipe da Laura tem determinado[5] essas características do sistema por grupo de interesse:

- **Para a administração hospitalar:**
  - reduz os custos gerais de hospitalização;
  - melhora a eficiência no rodízio de leitos (mais pacientes atendidos);
  - gera relatórios e mostra as tendências em tempo real;
  - promove uma transformação digital.
- **Para a equipe médica:**
  - prevê a deterioração do paciente por meio do uso de inteligência artificial;

5 Veja https://empowering-changemakers.eu/wp-content/uploads/2020/02/Projeto-Laura-BR.pdf

- permite intervenções mais eficazes com decisões baseadas em dados;
- gera relatórios e exibe tendências em tempo real, com informações clínicas agregadas (linha do tempo do paciente);
- alerta e ativa a equipe de resposta rápida.

- **Para a equipe de enfermagem:**
  - recebe as informações;
  - produz alertas precoces, para que a equipe assistencial possa atuar nos pacientes de maior risco;
  - reduz a sobrecarga de trabalho;
  - reduz a fadiga e o estado de alerta relacionados à sobrecarga de informações;
  - aumenta a eficiência operacional da equipe;
  - ajuda a equipe assistencial a priorizar suas tarefas (informações que fundamentam ações);
  - gera relatórios e exibe estatísticas em tempo real e virtualmente.

- **Para os pacientes:**
  - percebem um controle sistemático de seu estado;
  - percebem um ambiente hospitalar com tecnologia moderna e avançada.

MODELO ALGORÍTMICO

O sistema Laura de detecção de deterioração clínica usa um **algoritmo de Reforço de Gradiente ou *Gradient Boosting*.** Trata-se de uma técnica de aprendizado de máquina para problemas de regressão e classificação usada para prever eventos menos comuns (que correspondem a menos de 5% do conjunto de dados ou *dataset*). O sistema Laura possui um modelo de predição geral que foi treinado com 121.000 dados de atendimentos hospitalares únicos entre 2016 e 2019, provenientes de seis hospitais em diferentes regiões do Brasil (Rio Grande do Sul, Paraná e Minas Gerais).

Os **dados de treinamento** usados são seis:

1. Sinais vitais (temperatura, saturação de oxigênio, frequência respiratória, nível de glicose no sangue e pressão arterial).
2. Sexo biológico.
3. Idade.
4. Quarto.
5. Setor no qual o paciente está internado.
6. Duração da internação em dias.

O resultado a ser previsto é a **mortalidade hospitalar.**

Para criar os dados de treinamento, considera-se uma janela de 36 horas antes do resultado a ser previsto. A partir dessa janela, as últimas 12 horas são descartadas para evitar vieses no modelo. Assim, o modelo utiliza cinco coletas de sinais vitais em um período de 24 horas como dados de treinamento.

O modelo de aprendizado de máquina **adapta-se às necessidades e condições de cada centro.** Isso ocorre de duas maneiras: retreinando-se o modelo com dados locais, quando disponíveis, e negociando-se a sensibilidade do modelo com a equipe médica local. Para isso, o hospital ou centro clínico no qual o sistema Laura está implantado precisa dispor de um protocolo pré-determinado para o atendimento de pacientes com risco de deterioração clínica (incluindo variáveis relacionadas à alteração dos sinais vitais) que permita treinar o sistema em um ambiente real (Gonçalves et al., 2020). Com base nisso (que também inclui diretrizes sobre como os profissionais de saúde devem agir ao identificar alterações), o robô indica quais pacientes podem estar **em alto, médio e baixo riscos de deterioração clínica.**

A equipe da **Laura retreina um modelo específico para cada hospital** quando possui dados históricos suficientes (cinco anos) sobre seus pacientes. Se isso não for possível, **o modelo geral é usado.** Quer se utilize um modelo específico ou o modelo geral, sempre há um segundo processo de calibração do modelo com a equipe médica local. O sistema introduz testes durante um período, e a equipe médica local decide o limite a partir do qual uma probabilidade é considerada como alto risco.

ANÁLISE REALIZADA SOBRE O SISTEMA LAURA

O sistema foi avaliado em diferentes ocasiões por meio do estudo de suas diferentes versões. As análises concentraram-se em diferentes modelos algorítmicos e na eficiência de seus resultados. O trabalho de Kalil (2017) analisou retrospectivamente o impacto da implementação do sistema Laura no processo de identificação e manejo de pacientes com risco de sepse em uma unidade clínico-cirúrgica. Comparou as taxas do Tempo Médio de Atendimento (TMA), ou seja, o tempo médio de inserção de qualquer registro de dados no sistema de prontuários eletrônicos dos pacientes (evolução, dados vitais, prescrições, exames laboratoriais), calculado de forma autônoma pelo robô cognitivo seis meses antes e seis meses depois de sua implementação no sistema. O estudo não revelou mudanças significativas nessa taxa, mas destacou **o potencial do sistema de prever riscos** por meio de sua capacidade de mineração de dados.

Kalil et al. (2018: 312) confirmaram os resultados supramencionados. O objetivo desse estudo foi avaliar o impacto da implantação do sistema Laura nos processos relacionados à identificação e atendimento de pacientes com risco de sepse em uma unidade clínico-cirúrgica de um hospital privado de Curitiba/PR. Foram examinados os prontuários de 60 pacientes identificados com infecção e/ou sepse em um período de seis meses antes e depois da implementação da referida tecnologia no hospital, e foi avaliado o tempo médio de atendimento a partir da leitura autônoma do robô. As diferenças no tempo médio até a prescrição de antibióticos a partir do primeiro sinal de infecção identificado, com ou sem sepse, não foram estatisticamente significativas ($p = 0,85$). Em relação ao tempo médio de atendimento, observou-se uma **redução de 305 para 280 minutos** em relação aos seis meses anteriores e posteriores à implementação da tecnologia ($p = 0,02$).

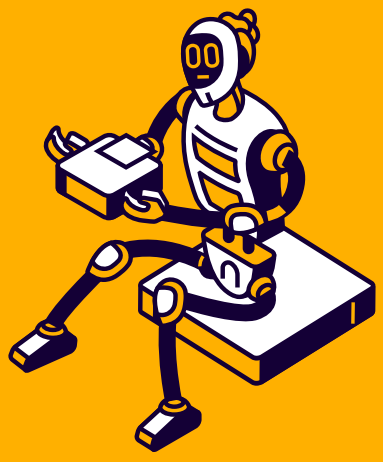
A pesquisa de Gonçalves et al. (2020) abordou a implementação da Laura nos aspectos relacionados à **interação entre equipe de enfermagem e tecnologia** em um hospital filantrópico durante o ano de 2018. Por meio de observação participante e entrevistas com

atores-chave, a administração e a operatividade do sistema em seu contexto de adoção foram analisadas qualitativamente. O sistema, ainda focado na identificação do risco de sepse, demonstrou ser **adotado de forma participativa** pela equipe de enfermagem, potencializando e agilizando a tomada de decisão na identificação precoce da sepse. Como resultado desse trabalho, recomendou-se que todos os casos de alerta fossem analisados e validados pelos profissionais de saúde do hospital.

O trabalho de Kobylarz et al. (2020) analisou 121.089 consultas médicas de seis hospitais diferentes e 7.540.389 pontos de dados. Os autores compararam os protocolos aplicados nos quartos para detectar a deterioração clínica com seis métodos de aprendizado de máquina escaláveis diferentes (três modelos clássicos de aprendizado de máquina, modelos baseados em taxas logísticas e probabilísticas, e três modelos conduzidos por gradientes, como o LightGBM). Os resultados mostraram **uma vantagem de AUC** (*Area Under the Receiver Operating Characteristic Curve*) de 25 pontos percentuais no melhor resultado do modelo de aprendizado de máquina em comparação com os protocolos atuais. Ao avaliar a hipótese da alternativa ou o complemento do sistema de IA de previsão de deterioração clínica nos quartos dos hospitais, o estudo revelou que o algoritmo que **apresenta melhores resultados é o LightGBM, com uma AUC de 0,961 e um F1 de 0,671.** Esse algoritmo apresenta uma maior precisão que o sistema MEWS, com pontuações de 0,697 AUC e 0,175 F1.

Assim, essas investigações indicam que o sistema Laura pode atingir uma certa precisão na identificação do risco de sepse ou deterioração clínica, bem como promover melhorias nos tempos de administração dos registros de dados hospitalares. Além disso, os estudos fornecem dados que permitiram ajustes no modelo. Por outro lado, revelam o potencial de implementação do sistema em ambientes hospitalares.

No entanto, esses estudos **não analisaram o impacto diferencial do sistema em diferentes grupos sociais** com base nas variáveis de precisão ou eficácia utilizadas. Esta auditoria concentrar-se-á, de forma complementar, no estudo do impacto diferencial do sistema por grupos protegidos, utilizando diferentes metodologias para identificar a eficiência da Laura.



2. ESTADO DA ARTE

2. ESTADO DA ARTE

O uso de sistemas baseados em **técnicas de aprendizado de máquina** está avançando rapidamente no setor de saúde (Sendak et al., 2020; Topol, 2019). Algumas das implementações desses sistemas têm demonstrado uma eficácia elevada na detecção e prognóstico de doenças, como, por exemplo, no caso da retinopatia diabética (Gulshan et al., 2016). No entanto, vale a pena considerar **possíveis efeitos indesejados da automatização** de testes clínicos e diagnósticos, que incluem violação da privacidade do paciente ou desumanização em seu tratamento (Ferryman e Pitcan, 2018; Madden, 2018). Esses são alguns dos motivos pelos quais a utilização desses sistemas é rigorosamente regulada por normas nacionais e internacionais, que os tornam suscetíveis a auditorias contínuas (Haupt, 2019; Price, 2017).

Esta seção analisará a literatura que trata de sistemas semelhantes à Laura e suas implicações, a fim de considerar o escopo geral e as principais limitações desses sistemas.

AUTOMAÇÃO DO RISCO DE DETERIORAÇÃO CLÍNICA

Há uma série de fatores **estruturais, tanto tecnológicos quanto relacionados aos recursos humanos**, que afetam significativamente as taxas de mortalidade hospitalar. Pacientes gravemente doentes geralmente apresentam alterações em seus sinais vitais durante um período antes de piorar. A falta de capacidade técnica e humana de detecção precoce dos pacientes que requerem tratamento prioritário demonstrou ter efeitos negativos nesse processo diagnóstico, resultando, em muitos casos, em um aumento das taxas de mortalidade e deterioração clínica (Pimentel et al., 2021; Goldstein et al., 2017).

Nas últimas décadas, **diferentes sistemas automatizados** foram desenvolvidos para identificar e notificar os **primeiros sinais de deterioração clínica** e fisiológica (Goldstein et al., 2017). A adoção de prontuários eletrônicos dos pacientes melhorou a disponibilidade de dados, que podem ser processados por técnicas de aprendizado de máquina para extrair informações que fundamentam decisões clínicas. Os modelos de aprendizado de máquina mais comumente usados para essa finalidade incluem regressão logística, métodos baseados em árvore, métodos baseados em kernel e redes neurais (Muralitharan et al., 2021). Um modelo algorítmico baseado no *Track and Trigger Scoring System (TT)* e que é fundamental no reconhecimento precoce é o ***Modified Early Warning Score*** **(MEWS).** Demonstrou-se que sua implementação em determinados contextos aprimora os mecanismos hospitalares de monitoramento dos sinais vitais dos pacientes. Da mesma forma, sugeriu-se que sirva de suporte para a equipe de enfermagem na identificação de pacientes em situação de risco, pois ajuda a assegurar sua situação clínica.

Outros modelos, desenvolvidos com a técnica de *deep learning* e que utilizam *Recurrent Neural Networks, the Long Short-Term Memory*, têm sido utilizados com sucesso para prever os sinais vitais do paciente e a posterior avaliação da gravidade de seu estado de saúde, utilizando Índices de Prognóstico (com uma precisão de 80%) (Bandeira da Silva et al., 2021). Com base nesse modelo, é possível prever futuros diagnósticos graves que não seriam identificados pela análise dos sinais vitais do paciente em sua situação atual. Outros sistemas de *deep learning* (redes neurais) são usados para detectar pacientes com risco de parada cardíaca, demonstrando, assim, alta sensibilidade e baixa taxa de alarmes falsos (Ueno et al., 2020). As diferentes finalidades clínicas desses sistemas no ambiente hospitalar e sua aplicação pelas equipes de enfermaria estão sendo ativamente estudadas. Em muitos casos, mostram que contribuem para um melhor acompanhamento dos sinais vitais do paciente e sua segurança clínica.

Os **dados de entrada** desses sistemas são múltiplos e dependem, entre outros fatores, da definição de deterioração clínica utilizada. Alguns têm demonstrado uma certa eficiência na identificação de riscos por meio da análise da linguagem natural, utilizando as anotações de enfermeiras e enfermeiros inclusas nos Prontuários Eletrônicos dos Pacientes (PEPs) (Zfania et al., 2020). As previsões baseadas apenas nos atributos coletados no momento da admissão hospitalar têm demonstrado ser altamente precisas na previsão do risco de readmissão nas Unidades de Tratamento Intensivo (Loreto et al., 2020), cujo estudo sugere que os "marcadores precoces" podem ser particularmente úteis para prever o risco de deterioração clínica em pacientes com alto risco de deterioração clínica após a alta da UTI.

No entanto, os diferentes algoritmos que estão sendo usados para detectar a deterioração clínica têm mostrado **diferentes graus de eficiência**, dependendo do contexto hospitalar e social, bem como dos preditores de risco usados. Por exemplo, dependendo de cada contexto social, terapêutico e organizacional de implementação, o algoritmo de *random forest* ou os algoritmos de regressão logística têm sido mais precisos na identificação do risco de deterioração clínica (Churpek et al., 2016). Além disso, esses autores demonstram a importância de uma boa calibração e do "*gradient boosting*" em preditores semelhantes.

Outras possíveis fontes de viés nesses sistemas estão relacionadas ao **desenho do modelo preditivo.** Um sistema de detecção de pacientes com deterioração clínica em internação que usa um esquema de pontuação foi modificado durante seu desenvolvimento e validação para reduzir os riscos de vieses **contra os pacientes mais velhos** (Pimentel et al., 2021). Embora, sob certas condições, os pacientes com mais de 80 anos tenham uma probabilidade decrescente de sofrer parada cardíaca ou ser transferidos para a UTI, os resultados dessa pesquisa mostraram uma variação mais ampla no risco geral previsto para pacientes com mais de 80 anos. A solução proposta para o problema foi incluir "uma ampla gama de fatores do paciente (comorbidades, fragilidade)" no modelo (Pimentel et al., 2021: 18).

Um estudo recente, que analisa os resultados fornecidos por vários preditores de risco no contexto hospitalar, evidencia a necessidade de considerar os **possíveis vieses integrados aos dados dos PEPs**, como a ausência ou a qualidade dos dados para determinadas variáveis (Goldstein et al. al. , 2017). Nesse sentido, é importante levar em consideração que a codificação eletrônica de determinados dados (por exemplo, relativos a decisões clínicas) varia entre hospitais e, em muitos casos, não é robusta o suficiente para ser incluída em um modelo generalizável (Pimentel et al., 2021).

Finalmente, o uso desses sistemas pode ser afetado pela **percepção de sua utilidade e vieses humanos** integrados ao uso dos sistemas ou aos protocolos de inserção de dados. Por exemplo, a literatura evidencia que a admissão na UTI varia de acordo com a experiência dos médicos e sua percepção dos benefícios e fatores organizacionais (por exemplo, a disponibilidade de leitos) (Green et al., 2018).

Dessa forma, o **modelo algorítmico e sua base teórica**, os **vieses históricos presentes nos dados** de entrada, os **processos de aprendizado** automático e o **viés no uso** são as principais fontes de discriminação algorítmica que também devem ser consideradas no estudo da Laura.

3. CONTEXTO SOCIAL

3. CONTEXTO SOCIAL

Esta seção descreve brevemente o **contexto social de implementação da Laura**, atualmente utilizada em centros clínicos e hospitalares de diferentes estados do sul do Brasil. Embora esta auditoria não tenha como objetivo contrastar empírica e comparativamente possíveis vieses do sistema Laura em uma ampla amostra de hospitais de todo o Brasil, esta seção busca estabelecer um quadro geral para sua análise e ilustrar as diferenças sociais estruturais que podem ser integradas ao sistema na forma de discriminações indesejadas.

A República Federativa do Brasil é formada por **26 estados e um Distrito Federal**, onde está localizada Brasília, sua capital. Os estados estão organizados em cinco regiões geográficas (Norte, Nordeste, Sudeste, Sul e Centro-Oeste), que apresentam importantes diferenças econômicas, culturais e demográficas. O país tem aproximadamente **209 milhões de habitantes** (em 2018)[6]. Embora a expectativa de vida tenha aumentado desde o Censo de 1999, as taxas de natalidade vêm diminuindo há décadas e caíram para menos de dois filhos por mulher[7]. Dessa forma, percebe-se um gradual envelhecimento populacional, embora continue sendo um país com uma população considerável abaixo dos 50 anos. Segundo o Censo de 2010, a população indígena brasileira é de 896.917 habitantes, o que equivale a 0,5% da população total do país[8].

A pobreza e a desigualdade social são significativas no país. A população pobre aproxima-se de 20% da população total, quando considerada a linha de pobreza para a classe de renda média-alta (13,8 reais em 2018) por dia/pessoa[9]. Em 2010, os estados das regiões Sudeste, Sul e Centro-Oeste apresentaram índices de desenvolvimento humano (IDH) altos ou muito altos (acima de 0,699), enquanto o **Nordeste e o Norte apresentaram índices de nível médio (0,600 e 0,699, respectivamente)**. Segundo dados da Organização das Nações Unidas, o país como um todo tem um IDH alto (entre 0,700 e 0,7999)[10]. Porém, enquanto a maioria dos estados do sul tem um PIB per capita acima de 10 mil dólares, (chegando a 13.299 dólares no caso de São Paulo), muitos dos estados do Norte estão abaixo de 7 mil dólares (IBGE, 2018).

O **Sistema Único de Saúde do Brasil** (SUS) é responsável por políticas que visam a garantir o acesso universal e integral aos serviços de saúde. Alguns de seus objetivos são a promoção da equidade, a gestão descentralizada e a participação social. A gestão do sistema é compartilhada pelos três âmbitos de governo: o Ministério da Saúde, no âmbito federal, e as secretarias de saúde estaduais e municipais, nos âmbitos regional e local. O sistema é financiado por impostos e contribuições nos âmbitos federal, estadual e municipal[11]. No entanto, o SUS tem cobertura limitada ou territorialmente desigual, o que pode afetar as taxas de deterioração clínica e o acompanhamento médico dos pacientes. Por exemplo, tem-se destacado que a disponibilidade de leitos livres em UTIs é um problema muito importante e generalizado no país (Cardoso et al., 2011).

6 Veja: World Bank data. https://datatopics.worldbank.org/world-development-indicators/

7 Veja: Organização Pan-Americana da Saúde, com base em dados do Departamento de Assuntos Econômicos e Sociais das Nações Unidas. Divisão de População. Nova York, 2015.

8 Veja: Instituto Brasileiro de Geografia e Estatística – IBGE. Nativos. Disponível em: http://indigenas.ibge.gov.br/graficos-e-tabelas-2.html

9 Veja: https://databank.worldbank.org/data/download/poverty/33EF03BB-9722-4AE2-ABC7-AA2972D68AFE/Global_POVEQ_BRA.pdf

10 No âmbito municipal, quase 80% da população residia em municípios com IDH baixo ou muito baixo em 1991; em 2010, entretanto, essa participação caiu para 11%. Veja: Programa das Nações Unidas para o Desenvolvimento – PNUD. Atlas. Disponível em http://www.atlasbrasil.org.br/2013/pt/o_atlas/idhm/

11 Segundo o Instituto Brasileiro de Geografia e Estatística, a despesa total com saúde em 2013 foi de 8% do PIB do país, com 3,6% representando despesas públicas.

As referidas variáveis sociodemográficas (pobreza, taxa de natalidade) e a abrangência do sistema nacional de saúde são alguns dos fatores que podem condicionar estruturalmente o risco de deterioração clínica das pessoas hospitalizadas nas diferentes cidades e povoados do país. Dadas as características do modelo da Laura, majoritariamente baseado em dados clínicos (temperatura, saturação de oxigênio, frequência respiratória, glicemia e pressão arterial), mas também em dados demográficos e hospitalares (sexo biológico, idade, quarto, setor onde o paciente está internado e duração da internação em dias), cabe indagar sobre os possíveis vieses **derivados de sua implementação em organizações e contextos sociais específicos.** A complexa relação entre esses fatores e as possíveis diferenças nos níveis de risco de deterioração clínica esperados para diferentes grupos sociais pode ser ilustrada, por exemplo, na distribuição territorial e na taxa de crescimento da prevalência e mortalidade por diabetes, que são maiores nas regiões Norte, Nordeste e Centro-Oeste do país (Duncan et al., 2020). Também existem outras variáveis grupais ou cruzadas que podem atuar como preditores de deterioração clínica. Por exemplo, um estudo com 271 crianças, realizado no Hospital Estadual da Criança da Bahia, revelou que o sexo masculino tem sido mais prevalente entre os menores que apresentam deterioração clínica (Miranda et al., 2020), o que confirma estudos que mostram que esse grupo prevalece nas internações em UTI e no que diz respeito às condições respiratórias (Batista et al, 2015; Time, 2007).

Com base no exposto acima, deve-se considerar que **existem preditores de risco de deterioração clínica que podem favorecer a discriminação algorítmica devido a vieses históricos** (dados reais que são afetados ou enraizados em questões de discriminação, legado ou políticas injustas). Dado que o sistema Laura foi implementado por diferentes instituições públicas e privadas no sul do país, esses elementos devem ser levados em consideração na análise dos vieses que resultam em um impacto diferencial em determinados grupos populacionais, quando podem transcender a capacidade de modelagem do risco estabelecida em cada centro com base em suas condições sociodemográficas e clínicas específicas.

4. ESTRATÉGIA METODOLÓGICA

4. ESTRATÉGIA METODOLÓGICA

JUSTIFICATIVA TEÓRICO-METODOLÓGICA

Um dos eixos fundamentais da auditoria algorítmica é a **identificação e análise da discriminação algorítmica.** Esta seção delineará esse conceito, fornecendo as principais definições usadas para identificar e analisar as diferentes formas de viés algorítmico injusto ou discriminatório.

Para contextualizar o viés algorítmico, é necessário, primeiro, distinguir entre diferentes formas de discriminação. Seguindo as definições de Lippert-Rasmussen (2013), a discriminação de gênero ocorre quando alguém trata a pessoa A pior do que trataria outra pessoa B, porque A tem algum atributo que B não tem. A **discriminação de grupo ocorre quando o referido atributo consiste simplesmente em pertencer a um grupo socialmente destacado**, ou seja, o pertencimento a esse grupo "é importante para a estrutura das interações sociais em uma ampla gama de contextos sociais" (Lippert-Rasmussen, 2013: 30). Ela também requer animosidade contra um grupo com base na crença de que as pessoas pertencentes a esse grupo são inferiores ou de que essas pessoas não devem se misturar com as outras.

Nesse sentido, **para ser considerado discriminatório, o viés deve envolver um ou mais dos chamados grupos protegidos**, que correspondem fundamentalmente aos atributos protegidos resumidos na Tabela 1. Essa síntese é baseada nos atributos protegidos abrangidos pela Lei de Igualdade do Reino Unido de 2010[12] (Seção 4) e pela Carta Europeia dos Direitos Fundamentais[13]. Vale destacar que esta lista não é exaustiva, pois pode ser adaptada ou modificada, dependendo do contexto[14]:

Tabela 1. Grupos e atributos (legalmente) protegidos

Grupos protegidos (não exaustivos)	Atributos protegidos
Crianças e idosos	Idade
Pessoas com deficiência (física e mental)	Incapacidade
Mulheres e transexuais	Gênero ou redesignação de gênero
Gestantes	Gravidez
Muçulmanos, judeus	Religião ou crença
Gays, lésbicas, bissexuais, intersexuais...	Orientação sexual
Pessoas de baixa renda/recursos escassos	Propriedade/recursos materiais

Fonte: elaboração própria.

12 Veja informações detalhadas em https://www.gov.uk/guidance/equality-act-2010-guidance

13 Legislação disponível em https://eur-lex.europa.eu/legal-content/ES/TXT/?uri=LEGISSUM%3Al33501

14 Os grupos desfavorecidos podem ser definidos em relação aos atributos mencionados no Artigo 21.º (Não discriminação) da Carta dos Direitos Fundamentais da União Europeia: "sexo (e gênero), raça, cor, origem étnica ou social, características genéticas, língua, religião ou crença, opinião política ou de qualquer outro tipo, pertencimento a uma minoria nacional, propriedade, nascimento, deficiência, idade ou orientação sexual". Esses grupos protegidos são definidos como indivíduos e grupos que compartilham uma ou mais das "características protegidas".

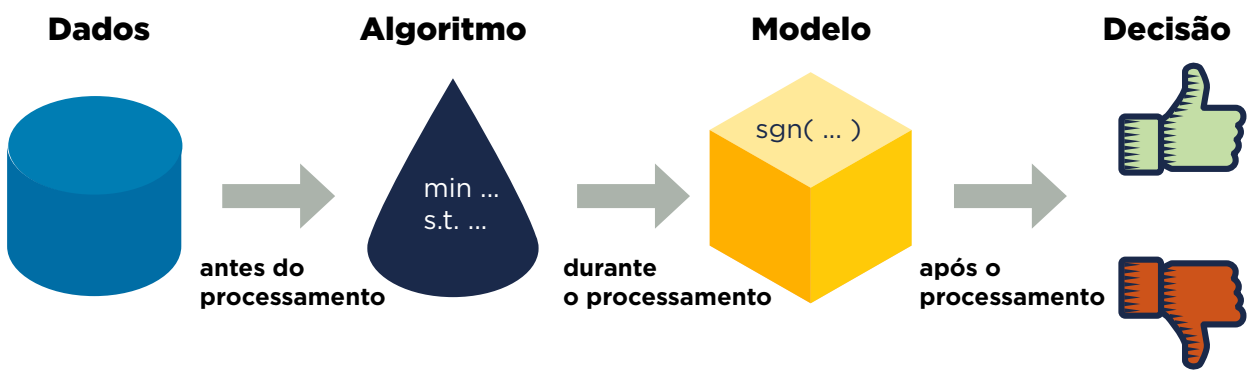
A discriminação estatística é uma **discriminação de grupo com base em um fato estatisticamente relevante.** Um exemplo clássico de discriminação estatística é não contratar uma mulher qualificada para um cargo de trabalho com a justificativa de que as mulheres têm uma probabilidade maior de tirar licença-maternidade. Por outro lado, a discriminação não estatística ocorre quando uma mulher não é contratada por dizer que pretende ter um filho e, consequentemente, tirar a licença-maternidade (Lippert-Rasmussen, 2013). Se for ignorada a animosidade correspondente aos humanos, mas não aos algoritmos, e qualquer característica usada no aprendizado de máquina for considerada estatisticamente relevante, pode-se dizer, então, que os algoritmos são capazes de discriminar (Castillo, 2018).

Deve-se levar em consideração que as definições dadas acima diferem das definições padrão de viés estatístico, que envolvem distorções de um cálculo estatístico resultante de amostras enviesadas ou estimativas cujo cálculo é incorreto em relação ao valor correto ou esperado de um parâmetro (Turney, 1996). Portanto, o viés estatístico não pode (sempre) ser um critério adequado de equidade algorítmica. Embora, pelo menos normativamente, o viés e a discriminação possam ser justos ou injustos, isso dependerá de como os resultados são interpretados social e eticamente. Seguindo a lógica acima, uma definição **mais precisa de viés algorítmico, ou discriminação algorítmica, implica a produção sistemática de resultados desvantajosos para grupos socialmente destacados, particularmente grupos desfavorecidos.** Esse viés está embutido nas propriedades matemáticas de um algoritmo.

O viés algorítmico foi dividido em dois tipos diferentes, dependendo do estágio do processo de aprendizado de máquina no qual ele ocorre (Danks e London, 2017). Em primeiro lugar, o viés algorítmico, assim como os modelos enviesados, pode ser **enviesado devido à coleta e uso de dados de treinamento enviesados** ao treinar ou modelar algoritmos durante os estágios iniciais de desenvolvimento (Cowgill, 2019). Em segundo lugar, o **viés pós-algorítmico ou de processamento** está relacionado à modelagem do sistema causada por suas interações com os usuários (processamento posterior no gráfico abaixo).

Nesse caso, o chamado tratamento díspar dos subgrupos pode basear-se em uma lógica aparentemente razoável, mas que ainda leva à discriminação (Barocas & Selbst, 2016). Portanto, a interpretação do usuário sobre o resultado do processamento algorítmico e o contexto social são fundamentais para avaliar se é justo ou injusto (Baeza-Yates, 2018). As diferentes fases durante as quais o viés algorítmico pode ocorrer, que são as mesmas fases em que o viés algorítmico pode ser mitigado, estão resumidas na imagem a seguir.

Figura 4. Estágios em que o viés algorítmico pode ser mitigado



Fonte: Hajian, S., Bonchi, F. e Castillo, C. (2016).

A JUSTIÇA ALGORÍTMICA

A definição e a sistematização da equidade algorítmica tornaram-se questões vitais para desenvolvedores e acadêmicos nesse campo (Gillen et al., 2018). De modo geral, a falta de equidade algorítmica pode ser definida como qualquer caso "em que os sistemas de AI/ML funcionam de maneira diferente para diferentes grupos de formas que podem ser consideradas indesejáveis" (Holstein et al., 2019: 3). Embora tenham sido desenvolvidos **métodos quantitativos para captar e medir o tratamento/impacto díspar nos grupos desfavorecidos**, essas técnicas não podem abordar o debate sobre quais grupos podem ser considerados desfavorecidos e o que pode ser considerado tratamento diferenciado em um determinado contexto sociocultural. De fato, a literatura evidencia a incompatibilidade habitual entre os modelos estatísticos de equidade e as interpretações feitas por usuários ou cidadãos (Binns, 2018; Kyung Lee, 2018). Esse debate manifesta-se **em múltiplas definições de equidade**, dificultando uma definição aceita unanimemente para uso por cientistas e engenheiros. Narayanan (2018), por exemplo, identificou 21 definições de justiça algorítmica.

Uma das definições mais importantes refere-se à **equidade de grupo, o que implica que o sistema algorítmico em vigor não deve tratar grupos sociais específicos de modo injusto**. Entre as medidas de equidade de grupo, destacam-se as três básicas descritas por Barocas e Hardt (2017): **independência** (também conhecida como paridade demográfica ou paridade estatística), **separação** (conhecida como probabilidades iguais ou prevenção de tratamento desigual) e **suficiência** (ou calibração), que são três das mais utilizadas na literatura.

- **Independência** significa que a probabilidade de atribuir um resultado independe do atributo protegido (por exemplo, na previsão de reincidência, quando a raça é o atributo protegido; isso implica que a fração de indivíduos atribuídos por um algoritmo à classe de alto risco será a mesma, independentemente da raça).

- **Separação** significa que a probabilidade de atribuir um resultado independe do atributo protegido, dado o resultado real (por exemplo, na previsão de reincidência, a fração de indivíduos atribuídos por um algoritmo à classe de alto risco será a mesma em todas as raças entre os indivíduos que não cometerem um novo crime no futuro).

- **Suficiênci**a significa que o resultado atribuído por um algoritmo não precisa ser combinado com atributos protegidos para que se obtenha uma previsão (por exemplo, na previsão de reincidência, uma determinada pontuação traduz-se na mesma probabilidade de cometer um crime, independentemente da raça).

Algumas dessas métricas de equidade de grupo podem ser incompatíveis entre si. Por exemplo, ao analisar sistemas de prevenção de reincidência, Chouldechova (2017) revelou que um instrumento que satisfaça a paridade preditiva não pode ter as mesmas taxas de falsos positivos e negativos entre todos os grupos, quando a prevalência de reincidência difere entre os referidos grupos.

Além disso, conforme indicado por Heidari et al. (2018), as noções estatísticas de equidade não garantem a equidade no âmbito individual. De fato, uma noção diferente de equidade algorítmica de grupo é a de **equidade algorítmica individual**, estabelecida pela primeira vez por Dwork et al. (2012) e que fala sobre um tratamento consistente dos indivíduos.

Para que um sistema seja justo do ponto de vista individual, **dois indivíduos (semelhantes no que diz respeito aos objetivos e ao modelo do algoritmo) devem receber resultados semelhantes.** Isso também ocorre se forem semelhantes quanto às suas características

na realidade, pois o modelo pode considerar iguais dois indivíduos que, na realidade, são diferentes, utilizando variáveis irrelevantes ou incorretas sobre eles. Portanto, esse modelo impõe restrições de tratamento a cada par de indivíduos (Kim et al., 2018). No entanto, como apontam Speicher et al. (2018), essas métricas não levam em consideração fatores contextuais mais amplos, como diferenças nas atividades anteriores realizadas por cada indivíduo ou o poder econômico ou social de cada um deles. Além disso, de acordo com Speicher et al. (2018), não existem mecanismos computacionais eficientes para integrar esses tipos de abordagem conceitual. Além disso, um sistema pode satisfazer o critério de equidade individual, mas gerar um resultado consistentemente adverso para um determinado grupo de indivíduos.

Assim, há um debate em andamento na literatura acadêmica sobre o desenvolvimento de métricas de equidade adaptadas a diferentes tipos de algoritmo e sistema. Além disso, observa-se uma **relação complexa entre equidade e eficiência**, pois, em alguns casos, a precisão preditiva pode ser prejudicada na tentativa de melhorar um sistema em termos de equidade (Narayanan, 2018). De fato, qualquer abordagem metodológica adotada com o objetivo de avaliar o viés deve combinar a análise de fatores específicos que determinam a equidade, de forma que permitam fazer uma contextualização sociológica e atingir os objetivos do processamento algorítmico. Para isso, deve-se levar em consideração o contexto social em que o sistema opera, dos pontos de vista quantitativo e qualitativo.

Nesse sentido, **este relatório segue um esquema proposto por Castillo (2018)**, segundo o qual os métodos algorítmicos que utilizam um critério para ordenar elementos, como pessoas, grupos ou semelhantes, devem ser capazes de alcançar a equidade em relação a estes fatores:

1. Uma presença suficiente de elementos do grupo protegido.
    a. Ausência de discriminação estatística (de grupo).
    b. Prevenção de danos atribuídos a um grupo.
2. Um tratamento consistente dos elementos de ambos os grupos.
    a. Ausência de discriminação individual.
3. Uma representação adequada dos grupos desfavorecidos.
    a. Prevenção de danos de representação em um grupo.

ABORDAGEM METODOLÓGICA DA LAURA

Levando em consideração as definições descritas acima, esta seção descreve brevemente as fases da auditoria algorítmica da Laura e a metodologia usada para avaliar a discriminação algorítmica.

FASES DA AUDITORIA ALGORÍTMICA

A auditoria é composta por quatro fases:

1. Estudos preliminares
Esta primeira fase dedicou-se, sobretudo, ao levantamento de informações básicas sobre as partes envolvidas na formulação, desenvolvimento e implementação do modelo, o modelo em si e sua integração às dinâmicas próprias das organizações que as partes envolvidas representam.

Para isso, as pessoas responsáveis pelo sistema Laura foram contatadas. Essas atividades permitiram coletar informações básicas sobre o sistema e as necessidades detectadas para seu desenvolvimento e implementação. Todas as informações necessárias para realizar a auditoria, bem como as decisões tomadas pela equipe de auditoria a partir de agora, estão refletidas neste e em outros documentos de trabalho internos.

2. Mapeamento da situação
Em segundo lugar, a equipe de auditoria estabeleceu como, quando, por que e para que desenvolveu e implementou esse algoritmo específico. Da mesma forma, examinou-se, por meio de um *Model Card* com dados sobre o sistema, se ele satisfazia uma lista de requisitos básicos para poder ser auditado e se as partes responsáveis por sua formulação, desenvolvimento e implementação estavam dispostas a fornecer as informações necessárias para sua realização.

3. Plano de análise
Esta fase consiste em definir e acordar com o cliente os termos (como e para quê) e os prazos estimados (quando) para o desenvolvimento da auditoria.

Com esse objetivo, foram realizadas diversas reuniões e trocas de informações com os responsáveis da *Eticas Research and Consulting* e da Laura. Com base nisso, a equipe de auditoria definiu um plano de análise (03/2021) a ser compartilhado e acordado com as partes envolvidas. Com base no acordo, elaborou-se e entregou-se uma proposta do plano de análise, e definiu-se a equipe de auditoria com conhecimento específico do sistema em questão.

4. Análise e relatório final
Esta fase focou-se na execução do Plano de Análise, dentro de uma certa margem de flexibilidade com relação ao que foi planejado, dependendo das circunstâncias do estudo. Neste caso, a análise correspondeu ao estipulado a seguir.

Em linhas gerais, a metodologia de auditoria algorítmica da *Eticas R&C* é realizada em duas partes complementares, cujo objetivo é entender a complexidade do modelo e suas possíveis implicações:

- Por um lado, um **estudo de caráter qualitativo**, com o objetivo de compreender as implicações do sistema e sua implementação, no contexto socioeconômico, técnico e organizacional no qual está inserido.
- Por outro, um **estudo do tipo quantitativo**, baseado em técnicas de análise estatística e ciência de dados, focado principalmente em detectar e recomendar medidas de correção para possíveis casos de imprecisão, discriminação, tratamento ou impacto diferencial, ou viés algorítmico causados pelo sistema. Vale destacar que, no caso da Laura, adotou-se uma política de proteção de dados que exigiu trabalhar com registros totalmente anonimizados, também no que diz respeito à confidencialidade do hospital analisado.

Nesta fase da auditoria, são realizadas as análises planejadas para a entrega final e são extraídos seus principais resultados.

APLICAÇÃO DO PLANO DE ANÁLISE ALGORÍTMICA

Foram realizados a análise do sistema automatizado e o estudo de seu viés e impacto diferencial por grupos, seguindo-se uma metodologia que vai do mapeamento do sistema algorítmico e seus dados de entrada/treinamento à aplicação de métricas destinadas a estabelecer diferenças estatísticas de acordo com os critérios de suficiência e independência mencionados na seção anterior.

Há quatro etapas principais na detecção de viés algorítmico:

(1) definição da atribuição de elementos a grupos;

(2) definição dos grupos protegidos;

(3) determinação de um conjunto de métricas destinadas a medir o viés; e

(4) medição e comparação entre os grupos.

A primeira etapa simplesmente **classifica os elementos de dados em grupos**, que podem ser sobrepostos (alocação *soft*) ou não sobrepostos (alocação *hard*). Essa sobreposição refere-se à convergência de mais de uma característica protegida a ser considerada; por exemplo, mulher de baixa renda. Na maioria dos casos, os dados refletirão os dados relativos a cada uma das pessoas, e, portanto, os grupos serão baseados em características individuais. Para criar esses grupos, pode-se utilizar qualquer característica atribuída a vários indivíduos, mas com atenção especial às características protegidas mencionadas acima. Esses agrupamentos são criados nos dados para avaliar até que ponto um algoritmo pode tratar ou afetar um grupo de maneira diferente de outro.

A segunda etapa determina **quais grupos foram definidos como protegidos**, o que significa que não devem ser prejudicados pela aplicação do algoritmo e que o impacto dos algoritmos sobre eles será monitorado de maneira especial. Em alguns casos, os grupos protegidos pertencem a categorias legalmente protegidas (por exemplo, pessoas com deficiência). Em outros casos, a definição do que constitui um grupo protegido refere-se a um compromisso que pode não ser juridicamente vinculante, como a intenção de aumentar a participação de mulheres ou minorias que podem estar sub-representadas em determinados cargos. Uma definição adicional de um grupo protegido pode basear-se na finalidade de uma tecnologia e, portanto, na conveniência do algoritmo. Por exemplo, se a intenção de um determinado algoritmo for aumentar a proteção de crianças de uma certa idade em um algoritmo para detectar chamadas telefônicas de denúncia de abuso doméstico, as crianças dessa idade constituirão um grupo protegido para o propósito da análise de viés algorítmico (Chouldechova et al, 2017).

A terceira etapa determina o **conjunto de métricas que serão usadas para a análise.** Em geral, essas métricas quantificam até que ponto um algoritmo trata as pessoas de maneira diferente (*disparate treatment*) e até que ponto um algoritmo tem um impacto diferente em pessoas diferentes (*disparate impact*). Existem várias definições de métricas, muitas vezes sobrepostas, que devem ser usadas para avaliar o viés algorítmico. No entanto, um certo grau de congruência deve ser mantido entre as definições em questão.

PROBLEMATIZAÇÃO, HIPÓTESE DE TRABALHO E MÉTRICAS

Apesar da extensão dos sistemas de inteligência artificial na área da saúde, suas implicações em termos de transparência e justiça algorítmica foram pouco estudadas (Sendak et al., 2020). Além disso, foram destacados os possíveis efeitos indesejados das novas tecnologias no campo da saúde (Ash et al., 2004).

Em primeiro lugar, revelou-se que certos fatores, preditores de risco ou bases teóricas para o diagnóstico clínico podem passar despercebidos ao basear a análise médica nos resultados oferecidos por sistemas baseados em inteligência artificial (Caruana et al., 2015).

Em segundo lugar, indicou-se que, em certos casos, a introdução de *machine learning* levou a uma **redução das capacidades constatadas do pessoal de saúde** na tomada de decisão (Hoff, 2011; Tsai et al., 2003).

Em terceiro lugar, e como já foi observado, deve-se levar em consideração que os algoritmos de aprendizado de máquina podem aprender a **prever riscos ou atribuir benefícios com base em informações enviesadas contra determinados grupos sociais.**

Foram identificados sistemas que oferecem resultados mais desvantajosos contra populações pobres ou não brancas ao avaliar variáveis como taxa de readmissão hospitalar ou mortalidade (Joynt Maddox et al., 2019; Lum e Isaac, 2016). Isso aconteceu em sistemas semelhantes ao sistema Laura. Por exemplo, uma auditoria de impacto diferencial em um algoritmo de medição de risco descobriu que, em uma determinada pontuação de risco, os pacientes negros estavam significativamente mais doentes que os pacientes brancos. Observou-se que remediar essa disparidade na medição de risco algorítmica **aumentaria a porcentagem de pacientes negros que recebem ajuda de 17,7 para 46,5%**.

A origem desse viés estava no fato de que o algoritmo previa os custos de atenção médica em vez da doença, mas a variável atenção médica era claramente desigual, o que afetava os pacientes negros (Obermeyer et al., 2019). Em outro caso, Di Martino et al. (2019) analisaram dois algoritmos (*Fetal Medicine Foundation e BCNatal*) para calcular o risco a priori de pré-eclâmpsia (com base no histórico médico dos fatores de risco) em cada indivíduo. Com uma taxa fixa de falsos positivos de 10%, os riscos estimados a priori pela *Fetal Medicine Foundation* e pelos algoritmos *BCNatal* em uma população italiana foram bastante semelhantes, e ambos eram confiáveis e consistentes. No entanto, os autores também verificaram que essa precisão é menor no caso das **gestantes sul-americanas**. Por isso, os fatores de aprendizagem que podem predispor um impacto diferencial por grupo desse tipo precisam ser analisados mais a fundo.

Com base nos dados de entrada do sistema de processamento automático no sistema Laura, nos tipos de viés algorítmico detectados em sistemas semelhantes e no escopo da auditoria, estimou-se que **devem ser avaliadas as categorias sexo biológico e idade na análise de impacto algorítmico diferencial**. Dessa forma, será possível estabelecer a capacidade do sistema de gerar um impacto específico nos pacientes em função das previsões de risco correspondentes a esses grupos. Com base nos resultados dessa análise quantitativa e do estudo de impacto social, serão propostas recomendações para monitoramento e estudo futuro.

Entre as diferentes métricas existentes, **as seguintes métricas são particularmente relevantes para o caso da Laura:**

M1. A precisão medida adequadamente para cada um dos grupos protegidos.

M2. As taxas de falso positivo e/ou falso negativo entre os grupos, que devem ser semelhantes, caso se queira afirmar que não há discriminação no funcionamento do algoritmo (ou seja, aplicar o critério de separação para evitar tratamentos desiguais, conforme descrito por Zafar et al., 2017). No caso em que receber um risco alto cria uma vantagem (em sistemas de assistência como a Laura), a taxa de falsos negativos para os grupos protegidos (que corresponde à porcentagem de indivíduos que sofrem deterioração clínica, mas que foram erroneamente categorizados como baixo risco) deve ser semelhante entre os diferentes grupos.

5. RESULTADOS DO ESTUDO OPERACIONAL E DE ACEITABILIDADE

5. RESULTADOS DO ESTUDO OPERACIONAL E DE ACEITABILIDADE

A equipe hospitalar vê os sistemas baseados em técnicas de aprendizado de máquina com desconfiança, devido à sua frequente falta de **explicabilidade e transparência.** Por esse motivo, certos sistemas algorítmicos nesse campo foram avaliados positivamente por sua forma de combinar alta precisão com uma apresentação dos resultados de modo a **facilitar sua interpretação** (Churpek et al., 2016).

Esta seção reflete uma primeira análise da implementação do sistema Laura e sua aceitabilidade, com base nos dados coletados até o momento sobre essas variáveis.

ALINHAMENTO DO SISTEMA NO CENTRO HOSPITALAR

Conforme já mencionado, durante o primeiro semestre de 2018, realizou-se um estudo de campo com observação direta e entrevistas semiestruturadas para conhecer o processo de trabalho dos enfermeiros que utilizaram a primeira versão do sistema Laura (Gonçalves et al., 2020). A análise apontou que a participação de enfermeiros começa na fase de desenvolvimento do sistema (fase de pré-implementação), quando compartilham conhecimentos científicos, teóricos e práticos de saúde, o que demonstrou ser fundamental para uma boa adoção tecnológica.

Nesse sentido, o processo de adoção da Laura é composto por diferentes fases de formação de recursos humanos e adaptação tecnológica. A empresa responsável por esse sistema possui uma equipe dedicada a orientar os centros hospitalares nos processos de **integração, adaptação e utilização da ferramenta.**

Na nossa entrevista com os responsáveis dessa organização (entrevista virtual, 18/3/2021), apontou-se que isso implica um apoio assistencial por parte de uma equipe com experiência clínica. Para tanto, são realizados diversos treinamentos com as equipes encarregadas da Laura. Essas equipes geralmente são compostas pelos diretores médicos e assistenciais, enfermeiros e médicos. A ferramenta é apresentada às equipes, e os diretores colaboram para o processo de implementação. Inicialmente, são apresentadas as fases da adoção tecnológica, e são explicados os protocolos de adaptação do sistema às necessidades e características do hospital. Em seguida, são realizadas duas fases de trabalho: alinhamento e implementação generalizada.

Figura 5. Processo de treinamento no sistema Laura

Fonte: Laura.

ALINHAMENTO

O processo de alinhamento engloba toda a integração do sistema, começando com uma ou duas áreas específicas do centro de atendimento. O processo é composto por uma parte técnica, dirigida pela equipe técnica da Laura, e um alinhamento assistencial, do qual participam especialistas da empresa, a equipe de enfermagem e a coordenação do centro hospitalar.

Esse processo consiste em:

A. **Alinhamento assistencial:** nesta fase, realiza-se um diagnóstico assistencial e clínico colaborativo, no qual são discutidos os processos clínicos do centro e suas especificidades de rotina (por exemplo, os pontos de entrada de informação clínica), que servirão de base para a implementação do sistema. O estabelecimento de um protocolo interno de atuação decorre do processo como um todo.

B. **Alinhamento técnico:** esta fase envolve a revisão dos bancos de dados do centro, seus sistemas e infraestrutura, e o ajuste do modelo algorítmico com base no estudo de sua eficiência aplicada à instituição. Nesta fase-piloto do sistema, também são realizados treinamentos no uso da plataforma e suas diferentes etapas de validação técnica.

IMPLEMENTAÇÃO GENERALIZADA DO SISTEMA

Em um segundo momento, como parte da expansão do uso da ferramenta para outras áreas do hospital, a empresa realiza novos treinamentos para equipes clínicas e novos eventos de validação técnica. Após essa fase de expansão, realizam-se treinamentos do pessoal, que, em muitos casos, inclui outras equipes hospitalares. Atualmente, o conjunto das fases de implantação ainda não está refletido em um Manual do Sistema para o hospital.

USABILIDADE

A Laura oferece vários produtos dentro do serviço dedicado à avaliação do risco de deterioração clínica: o relatório contínuo na **tela**, a **plataforma web**, que oferece dados em tempo real, e o **novo aplicativo** móvel, denominado Laura Assistant.

PLATAFORMA WEB

Como pode ser visto na imagem a seguir, a **plataforma web** oferece diversos instrumentos que vão além de informar o risco de deterioração clínica dos pacientes. Ela também relata o número de pacientes internados e monitorados (atenção ativa), os alertas existentes, ou seja, os pacientes com risco medido de deterioração clínica (alertas ativos), e o tempo médio de inserção dos dados clínicos sobre os pacientes [Tempo Médio de Inserção de Dados (TMID)], que inclui todos os dados (evolução, exames, medicamentos etc.) dos pacientes da instituição, não apenas a inserção de sinais vitais.

Figura 6. Interface do sistema Laura

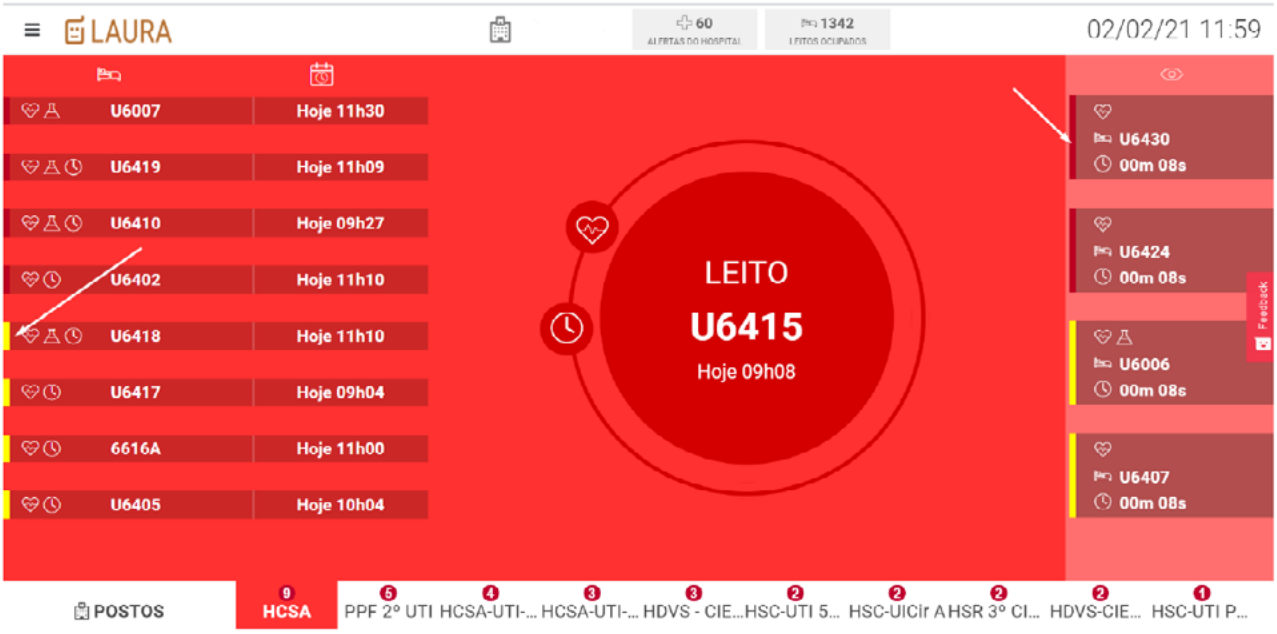
Fonte: Laura.

Além disso, o sistema permite que o pessoal autorizado busque informações específicas sobre cada paciente e dados estatísticos relevantes para o atendimento médico.

PAINEL DE EXIBIÇÃO

Em relação à **exibição dos painéis do sistema Laura**, o Painel de Controle Geral (veja a Figura 7) indica o estado clínico, o grau de criticidade e alertas dos pacientes. Pode ser visualizado em uma TV ou monitor. A gravidade da situação dos pacientes é expressa pelas cores vermelha (risco alto) e amarela (risco intermediário). A cor azul, por sua vez, indica um setor sem pacientes em alerta ou em observação.

Figura 7. Painel do sistema Laura



Fonte: Laura.

- O **paciente no centro do painel** corresponde ao paciente mais crítico do setor em um determinado momento, que, portanto, deve ser atendido primeiro.
- Enquanto os **pacientes com criticidade em amarelo** precisam ser reavaliados pela equipe a cada três horas, aqueles com criticidade em **vermelho** devem ser reavaliados pela equipe a cada hora. Esses tempos são configurados por cada hospital. O sistema registra o momento da reavaliação do paciente, inserindo apenas os sinais vitais no Prontuário Eletrônico do Paciente (PEP) da instituição.
- A parte inferior da tela apresenta **cada setor** do centro e o número de pacientes em alerta, segundo uma ordem vinculada ao número de pacientes em estado crítico. As informações de cada setor podem ser revisadas detalhadamente, clicando-se sobre o setor.
- O marcador que está ao lado de cada leito em alerta indica a **data e a hora da última atualização** do estado de alerta do paciente. Portanto, sinaliza quando o alerta foi iniciado ou atualizado em função dos novos dados inseridos.
- Com o objetivo de ilustrar o motivo do alerta, são utilizados três indicadores:
    - **um coração**, que indica "sinais vitais", quando o motivo do alerta está relacionado a alterações nesses sinais. Implica que a equipe deve verificar esses sinais, seguir os procedimentos clínicos adaptados à situação e inserir novos dados no sistema.
    - um **recipiente** para os "exames de laboratório", indicando que esses exames foram alterados, exigindo uma análise pela equipe médica.
    - um **relógio** para sinalizar um alerta de "TMID emergente", que indica que, além das alterações nos sinais vitais, o paciente não foi reavaliado no horário estabelecido para seu estado crítico. Isso exige uma reavaliação e uma reinserção de dados clínicos.

O canto esquerdo do painel de exibição mostra os pacientes em alerta ou que apresentam variações significativas em seus sinais vitais. Uma vez que os dados desses pacientes são atualizados, quando seu prontuário clínico é modificado, eles passam para o lado direito do painel, como pacientes em "observação". Para que o sistema classifique esses pacientes como reavaliados, devem ser incluídos pelo menos três sinais vitais (temperatura, frequência cardíaca e frequência respiratória) em seu prontuário médico.

Todos esses parâmetros, sistemas gráficos e dados estão descritos de forma clara e ilustrativa no **Manual de Instruções** do sistema Laura.

Os elementos que a literatura identifica como essenciais para a usabilidade dos sistemas de apresentação de informações de PEP, como **inteligibilidade das interfaces, estruturas não confusas do suporte de informação e iconografia coerente e intuitiva** (Raj et al., 2015), parecem ser devidamente considerados na tela do sistema Laura.

APLICATIVO LAURA ASSISTANT

O **aplicativo Laura Assistant** permite que os hospitais, com ou sem prontuário eletrônico do paciente, coletem facilmente sinais vitais de pacientes internados e processem esses dados em tempo real por meio do algoritmo do sistema Laura para detectar sepse e deterioração clínica. Todos os dados supramencionados, juntamente com outros, como resultados de análises ou notificações sobre os pacientes, são oferecidos por esse aplicativo.

Figura 8. Aplicativo do sistema Laura

Fonte: Laura.

ACEITABILIDADE NA LAURA

O estudo qualitativo realizado por Gonçalves et al. (2020) revelou uma aceitabilidade significativa do sistema pela equipe de enfermagem, ligada não apenas à utilidade do sistema, mas também à sua capacidade de transformar a dinâmica de trabalho ao fornecer informações em tempo real. Além disso, a Laura realizou pesquisas informais com os usuários finais do sistema que sugerem que o sistema **não aumenta sua sobrecarga de trabalho**, mas também que os médicos não **costumam interagir regular e ativamente** com a tela de informações (entrevista virtual, 18/3/2021). O atual desenvolvimento do aplicativo móvel, com uma interface que possui uma quantidade significativa de informações sobre a evolução dos pacientes, visa a solucionar essa deficiência e promover uma tomada de decisão informada.

Ainda **não foram realizados estudos sistemáticos sobre a usabilidade** do sistema. Os levantamentos realizados pela equipe da Laura revelam que os profissionais do hospital dão atenção especial às informações relacionadas aos pacientes em estado crítico. Esses pacientes são os que mais chamam a atenção da equipe médica para realizar uma avaliação imediata (entrevista virtual, 18/3/2021). Destacar sua condição de risco na exibição de dados é um elemento pessoal esperado. Nessa linha, o sistema foca-se no leito de maior risco e oferece diferentes medições, indicando os leitos sensíveis nas diferentes áreas do hospital (veja a Figura 7). As telas estão localizadas nessas diferentes áreas do hospital, para que a equipe de cada setor possa identificar os pacientes que possuem alertas no sistema Laura. A reação e atenção aos alertas varia de acordo com cada instituição e equipe médica (algumas indicaram um uso menor), mas, normalmente, a equipe deve **inserir informações no sistema** sobre a situação do paciente após sua avaliação. Um dos elementos que contribui para o bom funcionamento do sistema é a capacidade e a rapidez de cada equipe de integrar esses dados, avaliadas pelos estudos referidos na Seção 2. Isso pode ter um impacto significativo na **comunicação intra-hospitalar**, pois é necessário trabalhar com o painel, baseado em dados atualizados em tempo real. Da mesma forma, isso pode permitir uma boa transferência de informações entre os diferentes turnos de atendimento.

GERAÇÃO DO CONHECIMENTO CLÍNICO E TRANSPARÊNCIA

A equipe da Laura apontou que a existência de um mecanismo informático de sistematização e classificação sem aprendizado de máquina favoreceu um **maior conhecimento do desempenho hospitalar** e contribuiu para a **redução da mortalidade** associada ao controle de pacientes em estado crítico. A combinação de uma gestão organizada e aberta das informações, somada às notificações de atenção ou risco, com ou sem aprendizado de máquina, pode melhorar os processos de tomada de decisão. Além disso, o processo de pesquisa e implementação do sistema Laura levou à identificação de várias deficiências nos centros de atendimento clínico, como a falta de diretrizes e protocolos bem estruturados para detecção de sepse ou deterioração clínica. Assim, a relação da Laura com o hospital também faz parte de um processo de transferência de conhecimento que leva ao aprimoramento desses processos (entrevista virtual, 18/3/2021).

Da mesma forma, por meio da definição colaborativa dos indicadores de deterioração clínica, a Laura pode contribuir para a resolução de um dos problemas identificados pela literatura, que é a falta de consenso sobre como diagnosticar o surgimento dessa deterioração no setor pediátrico, onde os indicadores mais comumente usados incluem a necessidade de hospitalização ou transferência para a UTI (Bradman et al., 2014; Tucker et al., 2009; Miranda et al., 2020).

EXPLICABILIDADE ALGORÍTMICA

Deve-se levar em consideração que, segundo informações fornecidas pela equipe da Laura (entrevista virtual, 18/3/2021), os responsáveis técnicos e assistenciais dos hospitais são **informados sobre a precisão da Laura** com base na apresentação de sua taxa de falsos positivos, sua sensibilidade/ especificidade, *recall* e sua matriz de confusão após a primeira modelagem. Assim, procura-se não só adequar o sistema de inteligência artificial às características e necessidades de cada centro, como também conscientizar sobre o escopo do sistema e informar o processo de tomada de decisão, também no que diz respeito à versão implementada do modelo algorítmico. Por exemplo, como parte desse processo, as autoridades de saúde podem propor um equilíbrio específico entre a sensibilidade e a especificidade do sistema nesse processo. Além disso, a Laura disponibiliza aos centros relatórios semanais e mensais sobre seu funcionamento que demonstraram fornecer informações valiosas às equipes e promover ajustes nos protocolos médicos de atuação (entrevista virtual, 18/3/2021).

RESUMO DA ANÁLISE DE ACEITABILIDADE E RECOMENDAÇÕES ASSOCIADAS

Tabela 2. Síntese da análise de aceitabilidade e recomendações associadas

Dimensão	Análise	Recomendações
Usabilidade	Observam-se clareza e coerência no âmbito da inteligibilidade das interfaces, da estrutura do suporte de informação e da iconografia dos sistemas.	Recomenda-se a realização de pesquisas intra-hospitalares para estabelecer as limitações em relação a estas variáveis: inteligibilidade, clareza e coerência, e incorporar os resultados ao treinamento da equipe (incluindo materiais de apoio, como o Manual do Usuário) e à estrutura tecnológica.
Aceitabilidade	Tanto a literatura que aborda o sistema quanto as entrevistas sugerem uma boa recepção do sistema por parte dos usuários e usuárias. Além disso, a equipe do hospital indica que o sistema Laura contribui para melhorar o desempenho clínico dessas equipes. No entanto, o painel de informações também é utilizado de forma desigual, e o impacto do sistema na atualização dos dados dos pacientes nem sempre é significativo, como revela o trabalho de Kalil et al. (2018).	Recomenda-se testar a frequência de utilização do sistema em diferentes hospitais e áreas de atendimento clínico, tanto em termos de tempos quanto por meio de indicadores de desempenho na detecção e mitigação do risco de deterioração clínica. Nesse contexto, sugere-se também avaliar o impacto da utilização do Laura Assistant em termos da interação médico-máquina e da introdução de informação sobre o paciente (sinais vitais). Com base nisso, devem ser estabelecidos mecanismos, como treinamento ou aperfeiçoamento dos manuais, que permitam solucionar possíveis problemas de desempenho geral e dos diferentes setores.

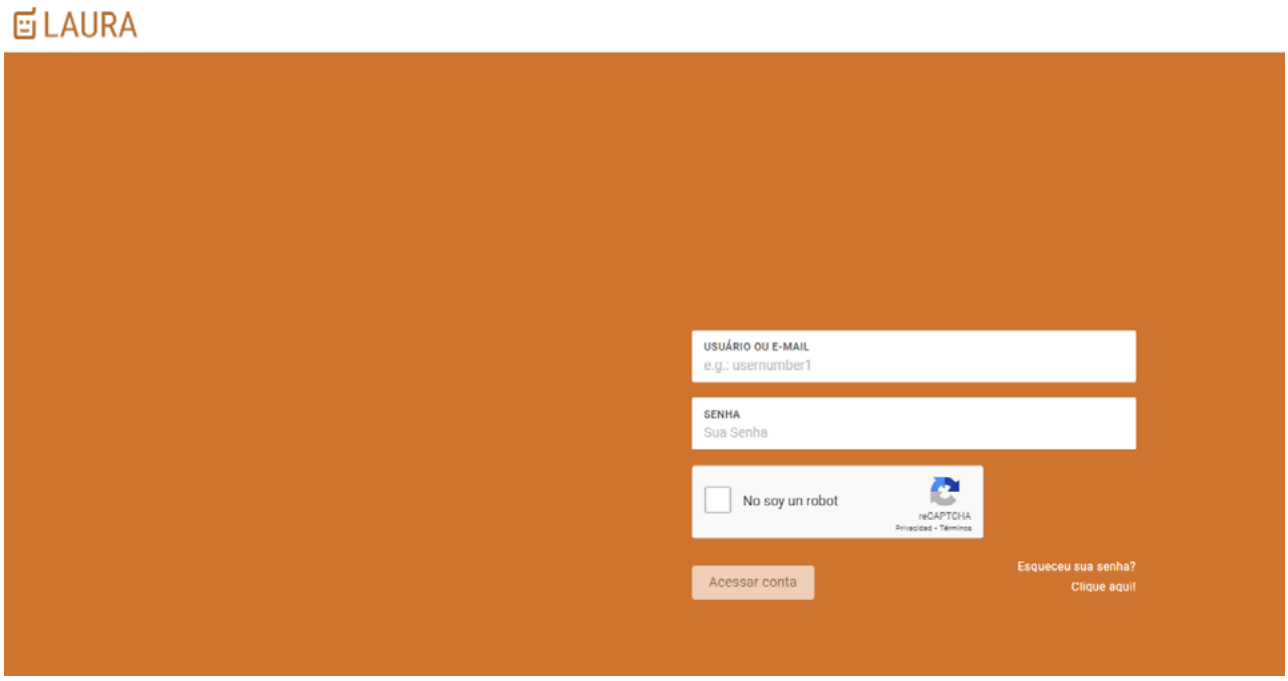
Geração de conhecimento e transparência	A utilização do sistema permitiu uma maior digitalização dos serviços hospitalares e a geração de conhecimento sobre o desempenho clínico. Essas informações servem tanto para melhorar o sistema de detecção de riscos quanto para o atendimento médico em casos de pacientes com baixo risco de deterioração clínica. Por outro lado, a geração de dados não se limita aos registros hospitalares, mas também inclui os resultados do processamento algorítmico. Esse conhecimento é compartilhado de forma aberta e colaborativa com a equipe técnica do hospital. No entanto, o conhecimento sobre o escopo e as limitações do modelo algorítmico (por exemplo, de precisão por grupos) não está sendo totalmente comunicado a toda a equipe durante o processo de alinhamento descrito na seção anterior.	Recomenda-se realizar validações regulares sobre a incidência da Laura na qualidade dos dados clínicos no registro digital do hospital. No que diz respeito à explicabilidade algorítmica, sugere-se a incorporação das informações de forma sistemática e compreensível para o público geral (Model Card - Mitchell et al., 2019) sobre o funcionamento do modelo algorítmico, incluindo: 1. Objetivos do sistema; 2. Dados; 3. Abordagem metodológica; 4. Descrição do algoritmo; e 5. Parâmetros de avaliação de desempenho e erros. Da mesma forma, essas informações devem ser comunicadas, como parte da política de privacidade do hospital, a todos os pacientes cujos dados serão medidos pelo sistema.

Fonte: elaboração própria.

RESULTADOS DA ANÁLISE DE GESTÃO/ADMINISTRAÇÃO DOS DADOS PESSOAIS

A Laura coleta e processa na nuvem os dados dos PEPs de cada hospital, principalmente em relação às variáveis de análise já mencionadas (entrevista virtual, 17/3/2021). Portanto, os hospitais estabelecem a conexão com o sistema via internet. As equipes técnica e hospitalar podem acessar o sistema em *https://laurabot.laura-br.com/#/access/login*, utilizando um **sistema de acesso** mediante inserção das credenciais de login (e-mail e senha de acesso).

Figura 9. Autenticação no sistema Laura



Fonte: Laura.

Como pode ser visto na figura anterior, o sistema Laura utiliza um CAPTCHA (*Completely Automated Public Turing test to tell Computers and Humans Apart*: teste de Turing público e automático para diferenciar computadores e seres humanos), que é uma medida de segurança para bloquear bots e evitar quebra de senha.

A **Política de Privacidade** do site da Laura (https://www.laura-br.com/politica-de-privacidade.html) incorpora os requisitos básicos de proteção de dados que correspondem às normas do Regulamento Geral de Proteção de Dados (RGPD). Isso inclui as finalidades do processamento, os dados pessoais envolvidos, os direitos ARCO e os detalhes de contato do responsável pela proteção de dados. No entanto, indica-se que os dados dos usuários que deixam comentários na web são **mantidos por tempo indeterminado para fins de identificação e filtragem**, o que não corresponde ao princípio da minimização de dados. Por outro lado, não são fornecidas informações detalhadas sobre terceiras partes envolvidas no processamento de dados pessoais.

No que diz respeito à gestão dos dados pessoais por parte do sistema Laura, os registros médicos são **processados de forma pseudonimizada** por meio da utilização de variáveis de deterioração clínica **associadas a um identificador não pessoal por paciente.** Esse identificador único seria o número do leito, que está associado ao número de registro (entrevista virtual, 18/3/2021). Da mesma forma, o registro é utilizado para encontrar o paciente na plataforma Laura.

Os **requisitos de segurança dos dados pessoais** são estabelecidos na Laura com base no contrato entre a empresa e o hospital contratante. As cláusulas de proteção de dados estão alinhadas às normas do RGPD e à Lei Geral de Proteção de Dados do Brasil (entrevista virtual, 18/3/2021). Além disso, as bases de conformidade nessa área estão sendo alinhadas aos requisitos do *Health Insurance Portability & Accountability Act*, a principal ferramenta legal de proteção de dados correspondente aos prestadores de serviços de saúde nos Estados Unidos.

Por fim, **15 dos 40 hospitais** onde o sistema está funcionando contam com **Comitês de Ética**, que podem avaliar a conformidade legal das investigações realizadas pela Laura com os dados inseridos dos pacientes.

ANÁLISE E RECOMENDAÇÕES

Observou-se uma tendência geral na direção da aceitabilidade do tratamento dos dados dos pacientes por meio do uso do PEP pela equipe do hospital (Alshahrani et al., 2021). No entanto, deve-se considerar que esses sistemas gerenciam dados sigilosos, como o registro médico do paciente, seu prontuário clínico, notas de progresso, medicamentos, sinais vitais, imunizações, dados laboratoriais, relatórios radiológicos e dados administrativos. De acordo com a nova Lei Geral de Proteção de Dados (LGPD) do Brasil (2020), que está alinhada aos requisitos do RGPD europeu, essa categoria de dados requer **precauções especiais em seu tratamento.** Essas medidas estão fundamentalmente ligadas à necessidade de **confidencialidade, integridade e disponibilidade das informações** em questão (Kuturura e Cilliers, 2016). Os principais riscos aos quais os sistemas de informação que gerem o PEP estão sujeitos em relação a esses requisitos incluem **ataques** de vírus ou **acessos não autorizados**, intencionais ou não, por parte da equipe hospitalar (Cilliers, 2017:3).

Na Laura, a capacidade de garantir a integridade e confidencialidade das informações está particularmente ligada à sua **estratégia de pseudonimização dos pacientes** submetidos à avaliação de risco e monitoramento. Se um conjunto de dados for anonimizado em alto nível, ou seja, se todos os dados pessoais forem removidos do prontuário médico recebido pela Laura, sua utilidade para terceiros diminuirá drasticamente. Da mesma forma, quanto mais útil é o conjunto de dados, menos anonimizado ele costuma ser (Ohm, 2009; Lubarsky, 2017).

Dependendo da técnica de anonimização, a compensação é diferente. Ou seja, cada técnica tem suas vantagens, deficiências e problemas associados. Nesse caso, optou-se pela utilização de um sistema de codificação para pseudonimizar o paciente e apresentar suas informações no sistema, utilizando-se o número do leito como identificador. A pseudonimização é uma forma de garantir a identificação e o vínculo contínuo com um ou mais conjuntos de dados sem identificar diretamente a pessoa. Normalmente, envolve a substituição de um valor, como um identificador pessoal, por outro valor. A pessoa cujo registro foi pseudonimizado permanecerá identificável devido à atribuição desse novo valor. Por exemplo, João Antunes passa a ser o usuário 3849562. Com esse sistema, uma pessoa que fez um exame pode consultar o resultado do exame no banco de dados com a identificação exclusiva que recebeu, sem que outras pessoas possam identificar uma pessoa específica. Trata-se de um método alternativo ao anonimato que às vezes é suficiente, dependendo dos dados e seus usos.

No entanto, se os quase-identificadores permanecerem dentro do conjunto de dados, o indivíduo ainda será reidentificável. A literatura estabeleceu que um pseudônimo não é útil para proteger a privacidade, se o mesmo pseudônimo único for usado continuamente em um ou mais conjuntos de dados (Lubarsky, 2017; *Article 29 Data Protection Working*

Party 29, 2014), principalmente se o número de atributos vinculados a um registro for alto e crescente (Barocas e Nissenbaum, 2014), como no caso do sistema Laura. As possibilidades de vinculação, singularização e inferência permanecem as mesmas entre um conjunto de dados pseudonimizados e o conjunto de dados original[15]. Dessa forma, o *Working Party* enfatizou fortemente em seu documento de opinião (2014) que um conjunto de dados pseudonimizados não é anonimizado nem atende aos padrões de anonimização. Porém, a pseudonimização pode ser usada em combinação com outras técnicas de anonimização, com o objetivo de anonimizar de forma robusta um conjunto de dados.

Além dessas questões, as medidas a serem consideradas no sistema Laura devem incluir:

- Monitoramento da segurança no **mecanismo de acesso ao sistema Laura.** A identidade da equipe que acessa o sistema deve ser verificada por meio de um mecanismo de autenticação seguro que se adapte aos privilégios de cada usuário (níveis de acesso) (Cilliers, 2017). Caso a biometria não seja utilizada, recomenda-se a utilização de protocolos de acesso que combinem senhas/códigos de identificação pessoal com alguma informação conhecida apenas pelo funcionário (por exemplo, crachá de identificação hospitalar).

- Para confirmar a **integridade das informações** transmitidas pelo sistema Laura, é possível implementar o *file hashing*, por meio da utilização de um algoritmo de filtragem que usa o valor dos bits do arquivo para confirmar sua qualidade e evitar que o documento seja comprometido (Laudon e Laudon, 2010). Classificar o nível de sigilo das informações, sujeitas a diferentes níveis de acesso, também pode contribuir para esse objetivo.

    - As informações também devem ser protegidas por firewalls e monitores de intrusão.

- Quanto à **transmissão e armazenamento**, como as informações são gerenciadas na nuvem, recomenda-se o uso de criptografia na comunicação com o sistema Laura. Além disso, sugere-se estabelecer algum mecanismo de *non-repudiation* para garantir que os dados sejam verificados no momento do recebimento por ambas as partes e que o hospital receba do sistema Laura a comprovação de identidade referente às informações enviadas, como as avaliações de risco realizadas (Maconachy e outros, 2001).

    - **Registro contínuo** de acessos (*logs*) ao sistema.

- A **formação dos funcionários** deve incorporar elementos sobre a proteção dos dados pessoais. Isso inclui tipos de dados gerenciados pela Laura, finalidades específicas, requisitos legais, medidas de proteção das informações pessoais e de transparência, e comunicação com os pacientes.

15 Isso ocorre principalmente se um algoritmo predeterminado for usado para pseudonimizar um conjunto de dados, conforme explicado por Lubarsky (2017).

Tabela 3. Síntese da análise de aceitabilidade e recomendações associadas

Variável	Análise	Recomendações
Política de privacidade	A política de privacidade da web é completa e está alinhada aos requisitos das normas internacionais referentes a esse campo.	Sugere-se rever o prazo de conservação dos dados pessoais e seguir o princípio da minimização dos dados, desde que sua eliminação não afete a qualidade do serviço nem a finalidade da tecnologia em questão.
Segurança de dados	O sistema Laura possui um sistema de nome, senha e Captcha para acesso aos dados.	Revisar o sistema de autenticação da identidade, incorporando códigos de acesso com informações conhecidas apenas pelo pessoal em questão. Integrar um mecanismo de registro e monitoramento dos acessos ao sistema, com monitoramento de eventuais acessos não autorizados. Garantir a boa qualidade e proteção dos dados por meio do file hashing e sistemas de proteção contra ataques. Garantir uma criptografia adequada na comunicação e armazenamento dos dados. Realizar treinamentos em proteção de dados para os membros da Laura e funcionários do hospital, abordando os requisitos básicos de proteção de dados sensíveis.
Pseudonimização e minimização de dados	O sistema de pseudonimização usado para apresentar dados de deterioração clínica reduz a exposição de dados pessoais. No entanto, trata-se de um sistema de codificação acompanhado por outros dados (como o número do leito), o que facilita a reidentificação da pessoa.	Recomenda-se revisar a política de pseudonimização para garantir o mais alto nível de confidencialidade e não reidentificação possível no âmbito da funcionalidade exigida.

Fonte: elaboração própria.

6. RESULTADOS DA ANÁLISE ALGORÍTMICA

6. RESULTADOS DA ANÁLISE ALGORÍTMICA

Esta seção apresenta uma síntese dos resultados da análise algorítmica com base na metodologia apresentada acima. O estudo visa a avaliar o **impacto diferencial do sistema Laura nos grupos afetados**, com foco nos atributos de sexo biológico e idade, bem como em suas interseções.

Antes de examinar os resultados da medição de impacto diferencial por grupos, descreveremos a estrutura dos dados de "entrada" do algoritmo, bem como os resultados observados ou reais a respeito da variável deterioração clínica na população estudada. Essa análise busca fundamentalmente entender a base histórica e real de aprendizado do algoritmo e contrastar suas previsões de risco.

A equipe da Laura disponibilizou um conjunto de dados (*dataset*) composto por **2.874 prontuários hospitalares** de um hospital do sul do Brasil onde o sistema Laura está em operação. Os registros foram processados durante o **ano de 2020.** As informações fornecidas compreendem um conjunto de dados com as seguintes informações: sexo, idade, setor do hospital, probabilidade do *outcome* (número entre 0 e 1), limite de *outcome* (número entre 0 e 1), previsão de *outcome* (número binário, sendo 0: alta hospitalar e 1: óbito) e *outcome* real (número binário, sendo 0: alta hospitalar, 1: óbito).

Para analisar o risco previsto pelo sistema Laura, o conjunto de dados foi formatado de acordo com os requisitos de equidade (*aequitas*); ou seja, as colunas *predict_outcome* são renomeadas como *score* e *real_outcome* como *label_value*, criando-se, assim, três grupos de análise com base nos atributos protegidos (sexo, idade e sexo + idade).
ESTRUTURA SOCIODEMOGRÁFICA
Os dados utilizados nesta auditoria são um conjunto composto pelos **2.874 registros hospitalares** mencionados acima. Como esta auditoria visa a estabelecer a eficiência do sistema na medição de risco por grupos de acordo com as **variáveis de sexo biológico e idade**, a população foi desagregada em grupos por sexo biológico e cruzada por faixas etárias.

No conjunto das 2.874 unidades, observou-se uma leve **predominância da população feminina** sobre a masculina no total de pacientes. Em segundo lugar, o grupo de pacientes do sexo feminino possui um número maior de registros para as faixas etárias mais jovens (18-29, 30-39, 40-49 e 50-59); portanto, há menos unidades nos grupos 60-69, 70-79 e 80+. Fica evidente que as faixas etárias **abaixo de 17 anos apresentam uma prevalência muito baixa**, não havendo nenhum registro para a faixa de 0 a 15 anos. Por fim, a tabela a seguir também apresenta os resultados de "alta hospitalar", que atingiram 85,4%, e de "óbito", representando 14,6% da população total.

Tabela 4. Representação demográfica de idade, sexo e resultado

Idade	#	%
0-15	0	-
16-17	7	0,2 %
18-29	119	4,1 %
30-39	241	8,4 %
40-49	435	15,2 %
50-59	567	19,7 %
60-69	775	27,0 %
70-79	528	18,4 %
80+	202	7,0 %

Sexo	#	%
Masculino	1.350	47 %
Femenino	1.524	53 %
Total	**2.874**	**100 %**

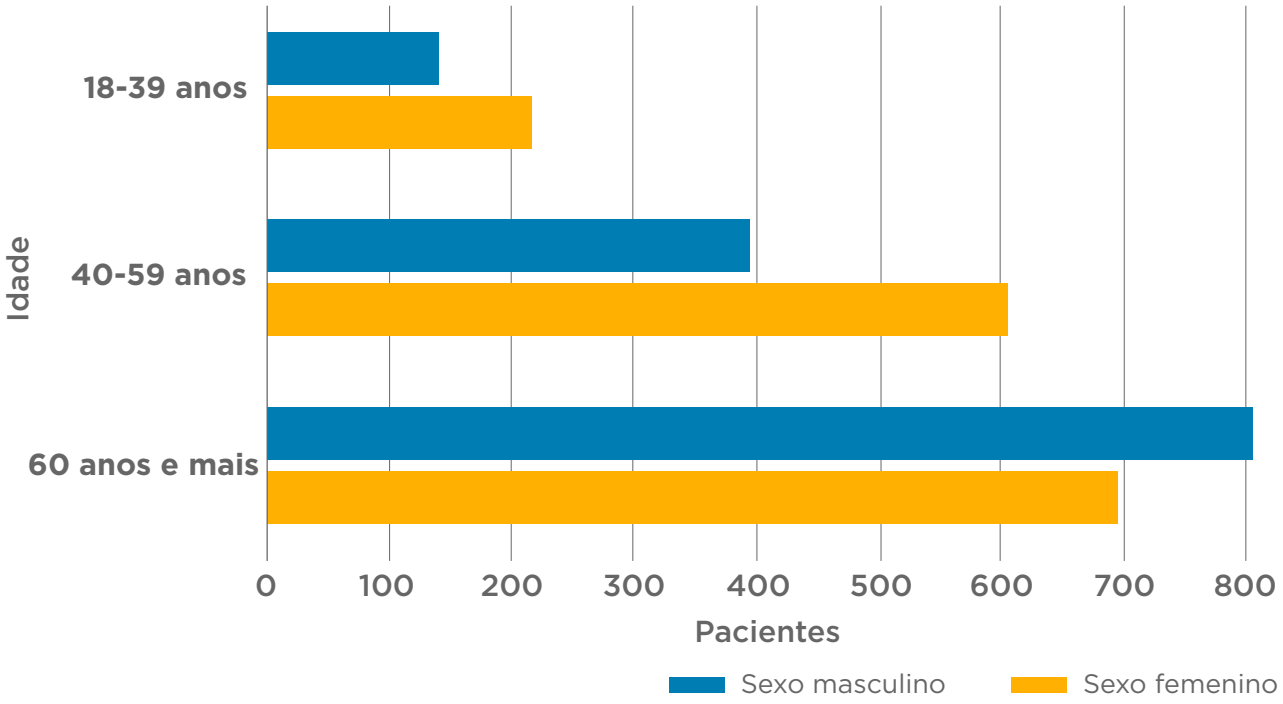
Resultado	#	%
Alta hospitalar	2.453	85,4 %
Óbito	421	14,6 %

Fonte: elaboração própria.

Na análise seguinte, optou-se por **eliminar os pacientes menores de idade**, pois seu número é muito pequeno (7 indivíduos no total), o que significa que o modelo não pode ser validado para essa população. Além disso, seu valor representa uma descontinuidade na curva de idade, o que pode indicar que existem circunstâncias externas que estão influenciando o aumento desse número.

Com base na nova distribuição de 18-39, 40-59, 60 ou mais, a **idade média dos pacientes** inclusos no grupo de dados examinados aproxima-se de 60 anos, embora seja um pouco maior para o sexo masculino. Essa estrutura dos dados de entrada é lógica, dado o risco crescente de deterioração clínica com o passar dos anos e a prevalência desse estado, contrastada pela literatura na população do sexo masculino.

Figura 10. Número de pacientes por faixa etária



ANÁLISE DE IMPACTO E TRATAMENTO DIFERENCIAL POR GRUPOS

Abaixo estão os resultados da análise do impacto diferencial do sistema Laura por grupos, de acordo com o sexo. Esse impacto é analisado por meio de três estratégias. Por um lado, compara-se o risco observado com o previsto por esse sistema para o mesmo grupo de pacientes. Por outro lado, são medidas e examinadas as taxas de falsos negativos (*False Negative Ratio*, FNR) e as taxas de verdadeiro positivo (*Positive Predictive Value*, PPV) para os mesmos grupos de idade e sexo. Por fim, a curva de calibração é identificada e analisada, a fim de estabelecer a relação entre valores de predição e percentual de positividade por grupos.

O QUE É ANALISADO COM A TAXA DE VERDADEIRO POSITIVO (PPV)?

O PPV é a taxa de verdadeiro positivo, ou seja, a **porcentagem de casos preditos corretamente entre todas as predições positivas feitas.** No sistema Laura, isso representa a proporção de pacientes que apresentaram deterioração clínica no período coletado, entre os pacientes para os quais foi previsto que isso aconteceria.

O **PPVd comparado por grupos (disparidade)** permite, portanto, projetar se o algoritmo prejudica determinados grupos já socialmente desfavorecidos. No contexto do algoritmo implementado pela Laura, isso significa que esse grupo de pacientes terá menos chances de ser atendido em uma situação de deterioração clínica.

Na análise intragrupo de disparidade do PPVd, considera-se como referência que o valor dos PPV esteja entre 80% e 125%, ou, em outras palavras, que o PPVd esteja entre 0,8 e 1,25.

O QUE É ANALISADO COM A TAXA DE FALSOS NEGATIVOS (FNR)?

A taxa de falsos negativos (*False Negative Ratio*, FNR) indica a probabilidade do modelo de prever que um paciente não está em risco de deterioração clínica, quando, de fato, ele está. Isso significaria que um paciente que precisa de ajuda não a receberia. **Um FNR mais alto indica uma maior probabilidade de subestimação do risco.** A Taxa de Falsos Positivos (*False Positive Ratio*, FPR) indica a probabilidade do modelo de prever que um paciente está em risco de deterioração clínica, quando, na verdade, não está, o que significaria que um paciente que não necessita ajuda a receberia, podendo deixar sem ajuda, como consequência, outro paciente mais necessitado.

O *False Negative Rate Disparity* (FNRd) é a proporção de pacientes com um resultado observado conhecido ("risco de deterioração clínica no período coletado") para a qual a previsão desse resultado é de "baixo risco" em relação a outros grupos. Nesse contexto específico, há interesse em ter uma taxa de falsos negativos baixa. Em outras palavras, gostaríamos de evitar casos de pacientes com maior risco de dano (no caso, "risco de deterioração clínica no período coletado"), mas que podem não receber os cuidados necessários, em parte porque o algoritmo prevê erroneamente o baixo risco.

Quanto à análise intragrupo da disparidade de FNRd, considera-se como referência que o valor dos FNR esteja entre 80% e 125%, ou, em outras palavras, que o FNRd esteja entre 0,8 e 1,25.

ANÁLISE DO IMPACTO DIFERENCIAL POR SEXO

RISCO OBSERVADO E RISCO PREVISTO POR SEXO

A tabela a seguir reflete a distribuição dos **resultados reais para os *outcomes* alta e óbito no conjunto de dados, sendo eliminado o grupo de menores de idade**, tanto em número total quanto em porcentagem de pacientes para cada grupo de acordo com o sexo. Como pode ser observado, 82,6% dos pacientes do sexo masculino internados e inclusos nesta auditoria tiveram alta no período analisado. Esse número é maior no caso do sexo feminino: 87,7% de altas no mesmo período. Em relação ao número e **taxa de óbitos**, é possível perceber que ela é **maior para o sexo masculino, com 17,4% do total.** Em contraste, essa taxa foi de **12,3% das 1.521 pacientes do sexo feminino.**

Tabela 5. Risco observado e risco previsto por sexo

	OBSERVADO				PREVISTO				
Sexo	**Alta #**	**Alta %**	**óbito #**	**óbito %**	**Alta #**	**Alta %**	**óbito #**	**óbito %**	**Total #**
Masculino	1.111	82,6 %	235	17,4 %	926	68,8 %	420	31,2 %	1.346
Femenino	1.335	87,7 %	186	12,3 %	1123	73,8 %	398	26,2 %	1.521

Fonte: elaboração própria.

Ao analisar os **resultados previstos por sexo**, observa-se que o sistema atribui 73,8% de altas ao grupo do sexo feminino sobre o total desse grupo e 68,8% ao masculino, ou seja, um número menor. Em sentido inverso, os óbitos são maiores nos homens, tanto em número quanto em porcentagem.

Embora os dados previstos sigam uma mesma tendência na distribuição do risco por sexo que nos resultados reais (atribuem um risco de morte menor ao grupo do sexo feminino), **deve-se levar em consideração que o risco previsto nesses resultados supera amplamente o identificado nos dados observados.** Para o risco de óbito no grupo do sexo masculino, essa distância é de **31,2% previsto versus 17,4% observado**, e, no caso do sexo feminino, **26,2% previsto versus 12,3% observado**. Assim, a previsão dobra e triplica a taxa de óbitos real, respectivamente.

PREDIÇÃO POSITIVA POR SEXO

Conforme mostrado a seguir, o algoritmo prediz adequadamente o risco de deterioração clínica em cerca de 50% das vezes para ambos os grupos nos casos identificados como de risco. Além disso, a percentagem de risco prevista corretamente (*true positive*) entre todas as previsões positivas feitas pelo robô Laura é **menor para o sexo feminino que para o sexo masculino**, embora essa diferença geral que afeta o grupo protegido (feminino) seja relativamente baixa.

Tabela 6. Previsão positiva por sexo

Sexo	**Verdadeiros Positivos**	**Falsos Positivos**	**PPV**	**PPV Disparity**
Masculino	223	197	53,1%	1,21
Femenino	175	223	44,0%	1,00

Fonte: elaboração própria.

TAXAS DE FALSOS NEGATIVOS POR SEXO

Como pode ser observado na Tabela 7, a variação das taxas de falsos negativos (FNR) é muito baixa, e a disparidade entre essas taxas por grupos dos sexos masculino e feminino é inferior a 15%. Isso implica que o sistema tende a subestimar com pouca frequência o risco de deterioração clínica e que a frequência dessa **subestimação é quase a mesma para ambos os grupos, embora maior para o sexo feminino**. Deve-se notar que este último grupo é maior em número total de unidades no conjunto de dados e também na porcentagem de altas, o que poderia explicar essa tendência.

Tabela 7. Falsos negativos por sexo

Sexo	Verdadeiros Positivos	Falsos Negativos	FNR	FNR Disparity
Masculino	223	12	5,1 %	0,86
Femenino	175	11	5,9 %	1,00

Fonte: elaboração própria.

ANÁLISE DE IMPACTO DIFERENCIAL POR IDADE

RISCO OBSERVADO E RISCO PREVISTO POR IDADE

Ao analisar os *outcomes* coletados pelo sistema Laura por faixas etárias, nota-se que as **maiores taxas de alta ocorrem nos grupos de 18 a 39 anos**, e que essa taxa diminui a cada faixa até os 60 anos ou mais. Por outro lado, as **maiores taxas de óbitos ocorrem nos grupos entre 40 anos ou mais**, concentrando-se no grupo de 60 anos ou mais. Em relação aos números totais, o grupo entre 60 e 69 anos acumula o maior número de pacientes com alta e com óbito.

Tabela 8. Risco observado e risco previsto por idade

	OBSERVADO				PREVISTO				
Idade	Alta #	Alta %	Óbito #	Óbito %	Alta #	Alta %	Óbito #	Óbito %	Total #
18-39	328	91,1 %	32	8,9 %	304	84,4 %	56	15,6 %	360
40-59	890	88,8 %	112	11,2 %	786	78,4 %	216	21,6 %	1.002
60+	1.228	81,6 %	277	18,4 %	959	63,7 %	546	36,3 %	1.505

Fonte: elaboração própria

Os dados previstos também refletem uma distribuição em que o **maior percentual de altas se condensa nos grupos de 18 a 39 anos**, e de óbitos, no de 60 anos ou mais. É possível notar como o percentual previsto de óbitos cresce nas previsões do sistema Laura após 50 anos, mas **praticamente dobra o percentual observado em todas as faixas.**

PREVISÃO POSITIVA POR IDADE

Ao analisar a proporção de pacientes com risco de deterioração clínica, por faixa etária e no período coletado, para a quais a previsão desse resultado entre os resultados positivos foi percentualmente maior, observa-se uma **taxa um pouco maior para o grupo entre 18 e 39 anos**, enquanto, para o resto dos grupos, as diferenças no PPV são inferiores a 5%, e a disparidade por grupos (PPVD) é muito baixa.

A maior taxa de positivos entre os pacientes com risco previsto, entre 18 e 39 anos, não se deve tanto a suas taxas de deterioração real, mas ao fato de ser o grupo com o menor número de pacientes.

Tabela 9. Previsão positiva por idade

Idade	Verdadeiros Positivos	Falsos Positivos	PPV	PPV Disparity
18-39	30	26	53,6 %	1,12
40-59	106	110	49,1 %	1,03
60+	262	284	48,0 %	1,00

Fonte: elaboração própria.

TAXAS DE FALSOS NEGATIVOS POR IDADE

Entre os pacientes com risco de deterioração clínica, a **Laura tende a subestimar esse risco com mais frequência nos pacientes entre 18 e 39 anos**. Esse grupo também coincide com aqueles com menos pacientes, o que pode indicar um viés derivado da composição dos dados de entrada. Em contrapartida, no caso da faixa etária de 60 anos ou mais, o risco de deterioração clínica tende a ser superestimado com uma frequência ligeiramente maior que nas demais faixas etárias.

Tabela 10. Falsos negativos por idade

Idade	Verdaderos Positivos	Falsos Negativos	FNR	FNR Disparity
18-39	30	2	6,2 %	1,15
40-59	106	6	5,4 %	0,98
60+	262	15	5,4 %	1,00

Fonte: elaboração própria.

ANÁLISE DE IMPACTO DIFERENCIAL CRUZADA POR GRUPO DE IDADE E SEXO

RISCO OBSERVADO E RISCO PREVISTO POR IDADE E SEXO

Como pode ser observado na Tabela 11, **o risco previsto de morte é maior que o risco observado**, diferença que aumenta com a idade dos pacientes. As altas observadas excedem as previstas em todos os casos, mas com uma diferença não significativa em termos de tratamento

diferenciado. No entanto, embora a taxa de risco de morte prevista pelo sistema Laura seja muito semelhante entre os sexos masculino e feminino para a faixa etária de 60 anos e mais, essa diferença tende a ampliar-se entre os dois grupos nas faixas etárias de menor idade.

Tabela 11. Risco observado e risco previsto por idade

		OBSERVADO				PREVISTO				
Edad	**Sexo**	**Alta #**	**Alta %**	**Óbito #**	**Óbito %**	**Alta #**	**Alta %**	**Óbito #**	**Óbito %**	**Total #**
18-39	M	123	86,0 %	20	14,0 %	115	80,4 %	28	19,6 %	143
18-39	F	205	**94,5 %**	12	5,50 %	189	87,1 %	28	12,9 %	217
40-59	M	336	84,8 %	60	15,2 %	294	74,2 %	102	25,8 %	396
40-59	F	554	91,4 %	52	8,6 %	492	81,2 %	114	18,8 %	606
60+	M	652	80,8 %	155	**19,2 %**	517	64,1 %	290	35,9 %	807
60+	F	576	82,5 %	122	17,5 %	442	63,3 %	256	**36,7 %**	698

Fonte: elaboração própria.

PREDIÇÃO POSITIVA POR IDADE E SEXO

A Tabela 12 mostra a proporção de casos de risco positivo identificados pelo sistema Laura no conjunto de casos positivos (PPV). Em linha com os resultados apresentados acima, o PPV é maior **para o grupo do sexo masculino entre 18 e 39 anos que para o grupo feminino na mesma faixa etária**. Isso é evidenciado pelo número de falsos positivos por faixa etária, no qual o sistema apresenta uma taxa PP maior para o grupo protegido (feminino), bem como pela taxa de verdadeiros positivos, que é muito alta para o sexo masculino e quase 30% menor para as mulheres. Em contraste, essa disparidade diminui consecutivamente por faixa etária.

Tabela 12. Previsão positiva por idade e sexo

Idade	**Sexo**	**Verdaderos Positivos**	**Flasos Positivos**	**PPV**	**PPV Disparity**
18-39 anos	Masculino	19	9	67,9 %	**1,35**
18-39 anos	Femenino	11	17	39,3 %	**0,78**
40-59 anos	Masculino	58	44	56,9 %	1,13
40-59 anos	Femenino	48	66	42,1 %	0,84
60 o mias	Masculino	146	144	50,3 %	1,00
60 o mais	Femenino	116	140	45,3 %	0,90

Fonte: elaboração própria.

A **diferença de 0,78 a 1,35 no PPVd** implica que o risco de deterioração clínica em um determinado número de pacientes do sexo feminino entre 18 e 39 anos de idade pode estar sendo subestimado de forma frequente e diferenciada. Isso pode ser explicado por três fatores. Primeiro, pelo **número de pacientes** do sexo feminino (217), que é superior ao de pacientes do sexo masculino (143) nessa faixa etária. No entanto, deve-se notar que essa diferença é maior para o grupo 40-59 (396 homens e 606 mulheres), mas a disparidade na taxa de verdadeiros positivos é reduzida. Em segundo lugar, outra explicação é o **alto índice de altas apresentado pelo sexo feminino nesse grupo (95%)**. Em terceiro lugar, pode ser vinculado aos **dados**

clínicos processados pelo algoritmo como preditores de risco (saturação de oxigênio, frequência respiratória, glicemia ou pressão arterial), que podem refletir melhores condições clínicas para o sexo feminino nessa faixa etária de maneira estatisticamente significativa.

TAXAS DE FALSOS NEGATIVOS POR IDADE E SEXO

Finalmente, de acordo com o exposto anteriormente, a Tabela 13 mostra como o sistema Laura tende a **subestimar o risco de deterioração clínica com mais frequência em pacientes do sexo feminino entre 18 e 59 anos** que em pacientes do sexo masculino. Notavelmente, essa diferença também é observada para mulheres na faixa etária de 40 a 59 anos. Isso coloca em questão uma explicação do viés com base no número de pacientes processados. Da mesma forma, fragiliza a explicação das diferenças observadas entre os grupos dos sexos masculino e feminino na faixa etária de 18 a 39 anos, devido aos melhores dados clínicos observados no sexo feminino.

Tabela 13. Falsos negativos por idade e sexo

Edad	Sexo	Verdaderos positivos	Falsos Negativos	FNR	FNR Disparity
18-39 años	Masculino	19	1	5,0 %	0,86
18-39 años	Femenino	11	1	8,3 %	1,43
40-59 años	Masculino	58	2	3,3 %	0,57
40-59 años	Femenino	48	4	7,7 %	1,32
60 o más	Masculino	146	9	5,8 %	1,00
60 o más	Femenino	116	6	4,9 %	0,85

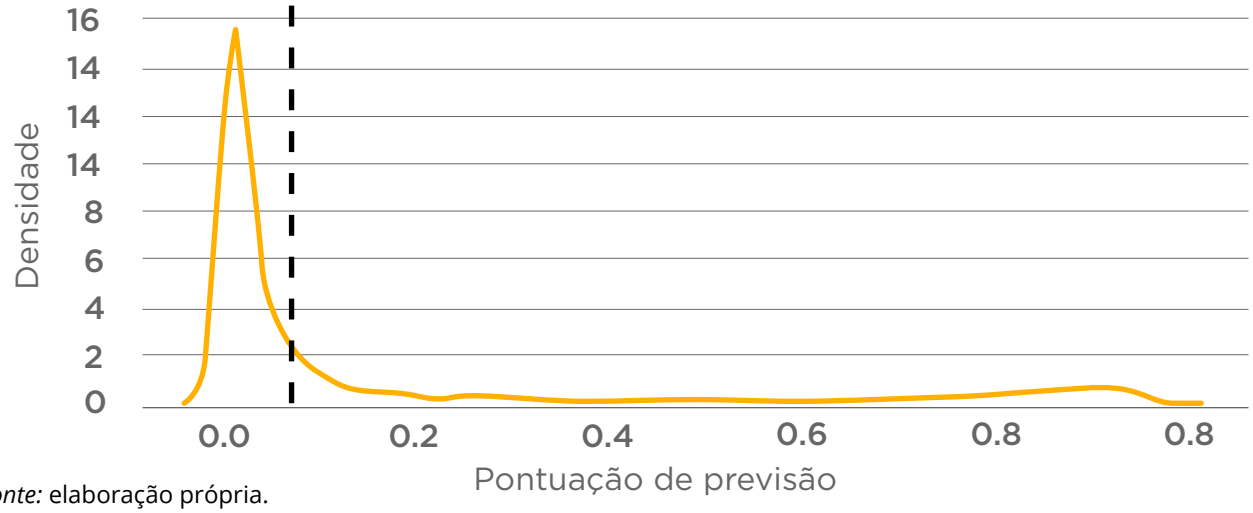
Fonte: elaboração própria.

ANÁLISE DA FUNÇÃO DE *SCORING*

A análise de calibração é baseada em uma definição inicial sobre a função de *scoring*. Trata-se de uma distribuição bimodal, entre 0 e 0,3, e entre 0,7 e 1,0. O corte de decisão em 0,069 (linha vertical na Figura 11) **é um corte arbitrário que obedece mais às características da equipe** e sua capacidade de resposta que a um risco extremo ou de morte.

O paciente que atinge uma pontuação de risco de 0,068 tem um risco semelhante a alguém que atinge uma pontuação de 0,070, embora sua interpretação binária seja muito diferente. Essa característica deve ser explicada à equipe médica que utiliza a Laura.

Figura 11. Curva de densidade de probabilidade



Fonte: elaboração própria.

ANÁLISE DA CALIBRAÇÃO

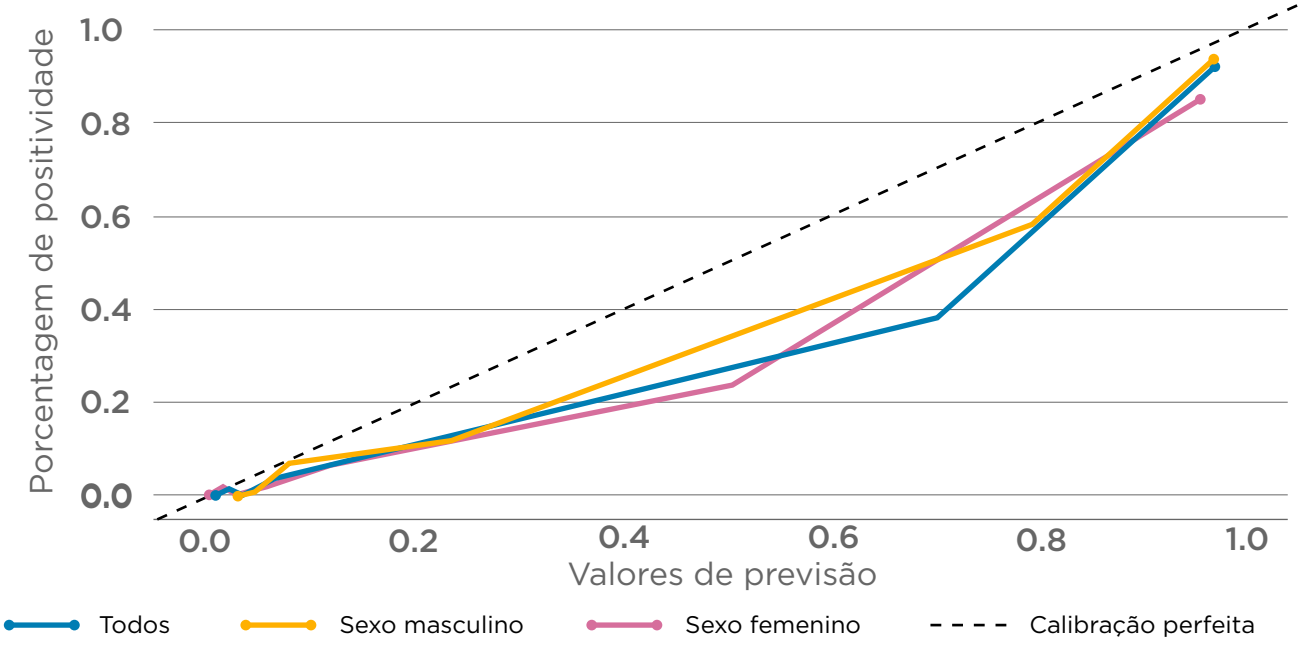
Uma vez definida a curva de densidade de probabilidade, é realizada uma **análise das curvas de calibração**, que permite avaliar a probabilidade de morte em diferentes grupos. Em particular, a curva de calibração indica:

- x = pontuação
- y = probabilidade de morte para pessoas com essa pontuação (número de pessoas que morrem dividido pelo número total de pessoas, dado um intervalo de pontuações).

Foram calculados decis sobre a pontuação para calcular seus intervalos, e esses limites foram considerados como o ponto de medição. **A "curva" resultante deve ser uma linha reta e igual para grupos diferentes** (M, F, M + *Young*, M + *Old*, F + *Young*, F + *Old*).

Como pode ser visto na Figura 12, a curva de calibração está mais ajustada à calibração perfeita nos casos de pontuações mais baixas e mais altas. A **calibração é perdida perto do ponto médio**, onde também são **observadas diferenças maiores entre homens e mulheres**. Em todos os casos, a pontuação parece superestimar o risco, ou seja, uma pontuação de 0,05, por exemplo, corresponde a um risco inferior a 5%, independentemente do sexo. Recomenda-se buscar uma melhor calibração do modelo, principalmente no que tange ao valor utilizado como ponto de corte.

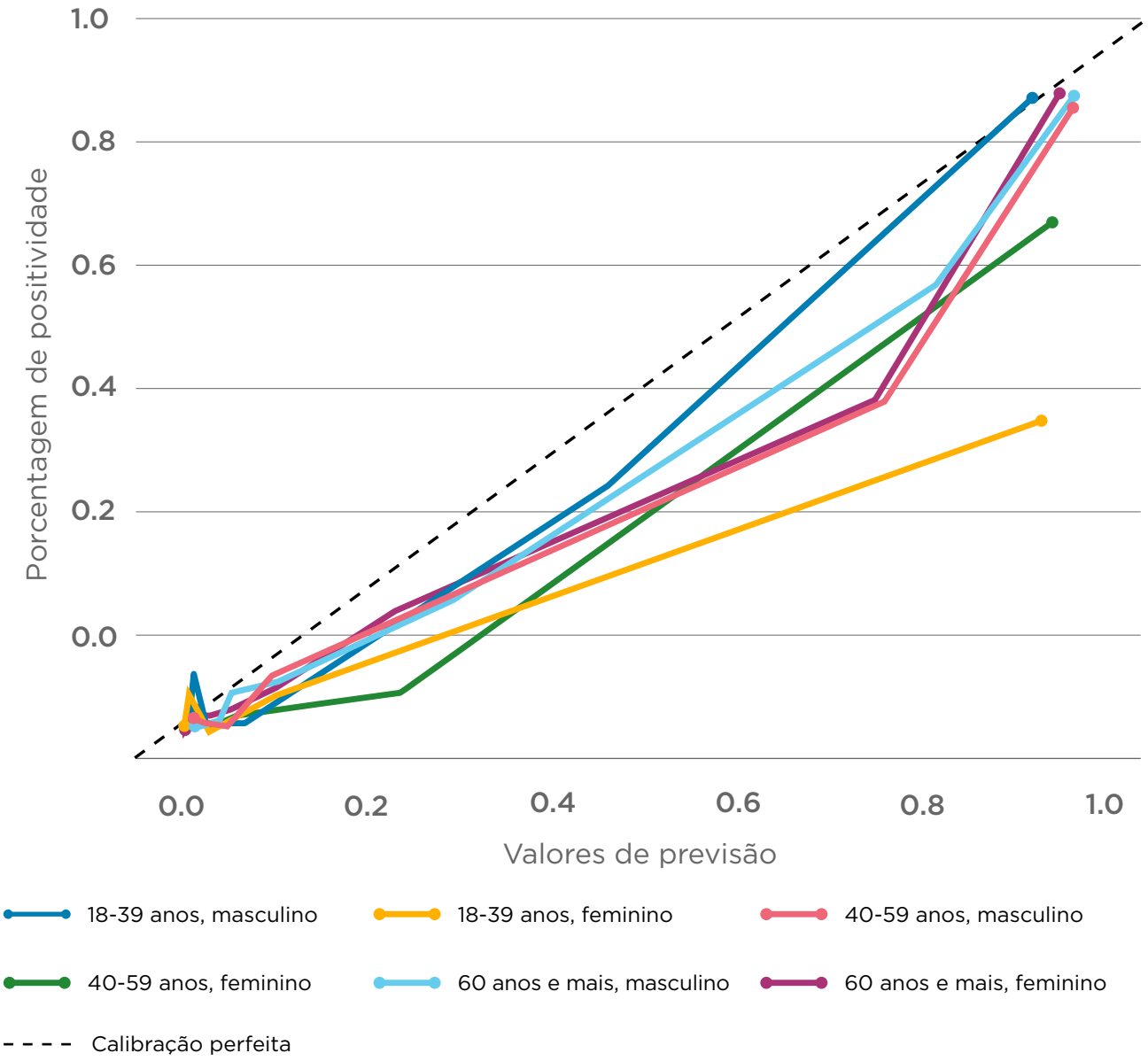
Figura 12. Curva de calibração por sexo



Fonte: elaboração própria.

Ao analisar essas curvas por faixas etárias, **observam-se importantes diferenças grupais** relacionadas às médias de risco predito identificadas acima. Por um lado, o grupo de mulheres entre **18 e 39 anos apresenta um distanciamento crescente no que diz respeito à probabilidade de mortes (descalibração), à medida que os valores de previsão aumentam**. O grupo mais bem calibrado é o do sexo masculino entre 18-39 anos, apesar de compartilhar uma taxa de mortalidade observada semelhante ao sexo feminino (14% e 15%, respectivamente) e incluir um número menor de pacientes. Por outro lado, enquanto os grupos dos sexos masculino e feminino apresentam a mesma curva para os grupos masculino 40-59 e feminino 60 ou mais, o grupo feminino entre 40-59 anos também apresenta uma queda na positividade a partir do valor de previsão de 0,2, que é sustentado continuamente ao longo da curva.

Figura 11. Curva de calibração por idade e sexo



Fuente: Elaboración propia.

CONCLUSÕES E RECOMENDAÇÕES

A auditoria do sistema Laura abordou vários aspectos de sua concepção e funcionamento com o objetivo de estabelecer uma aproximação de suas qualidades em termos de **aceitabilidade, usabilidade, proteção de dados e justiça algorítmica.** Para esse fim, estabeleceu-se, primeiro, o estado da arte teórico em torno dos sistemas de detecção automática de risco de deterioração clínica e a composição do quadro social geral para sua implementação. Em seguida, uma série de estratégias metodológicas e técnicas de coleta de dados foram aplicadas, visando a coletar dados qualitativos e quantitativos para realizar a análise. Isso incluiu entrevistas semiestruturadas com desenvolvedores e equipes encarregados da integração do sistema no ambiente hospitalar, bem como a medição de viés algorítmico de grupo com base nos falsos negativos e nos verdadeiros positivos por grupos.

Deve-se notar que a **análise de impacto social** foi projetada para fornecer uma visão geral dos aspectos sociais que poderiam limitar o funcionamento do sistema. Nessa linha, a auditoria algorítmica foi concebida para identificar **evidências indiretas de viés** a partir do estudo de duas variáveis sociodemográficas fundamentais para o sistema: sexo biológico e idade. Esse desenho metodológico, limitado à análise de fatores de viés específicos, é explicado pelo escopo da auditoria pactuada com o BID e refletida no Plano de Análise compartilhado como sistema Laura e adaptado à atual situação de pandemia, que limitou o trabalho de campo.

Em termos de impacto social, com base nas informações coletadas, nota-se uma **inteligibilidade significativa** por parte da comunicação in situ do sistema, nos âmbitos tanto da organização lógica das informações como da composição iconográfica. Em relação à aceitabilidade do Laura pelos usuários finais, foram identificadas diferentes fontes que indicam uma **aceitação tecnológica importante,** tanto pela **facilidade de uso** quanto pela construção e transmissão do conhecimento clínico. No entanto, essa boa recepção geral tem sido **atenuada pelo uso relativo do sistema** (Kalil et al., 2018), limitação que busca ser superada por meio do desenvolvimento do Laura Assistant.

Além disso, foram identificadas **limitações** na transmissão dos conhecimentos e informações aos usuários sobre o alcance e as limitações do modelo algorítmico, **particularmente no que diz respeito à precisão por grupos**, que poderiam ser comunicados de forma mais específica. Além disso, essa questão pode ser abordada com as pessoas que participarem da formação oferecida durante o processo de alinhamento da Laura.

Em termos de **proteção de dados**, foram identificados uma política de privacidade completa e bem estruturada, um padrão de acesso e autenticação do sistema com mecanismos básicos de segurança, e uma política de pseudonimização na transmissão aberta que seguem os princípios de privacidade em sua concepção. No entanto, sugeriu-se que esses padrões fossem revistos para confirmar sua proporcionalidade em relação à sensibilidade e ao volume de dados pessoais tratados.

RECOMENDAÇÕES DA ANÁLISE QUALITATIVA

I. Realizar **pesquisas inter-hospitalares** para estabelecer as limitações em relação às variáveis **inteligibilidade, clareza e coerência**. Incorporar os resultados dessas pesquisas na formação da equipe (incluindo materiais de suporte, como o Manual do Usuário) e na concepção da tecnologia.

II. Testar a **frequência de utilização do sistema** em diferentes hospitais e áreas de atendimento clínico, tanto em termos de tempos quanto por meio de indicadores de desempenho na detecção e mitigação do risco de deterioração clínica.

   A. Nesse contexto, sugere-se, também, avaliar o impacto da utilização do Laura Assistant em termos da interação médico-máquina e da inserção de informações sobre o paciente (sinais vitais). Com base nisso, devem ser estabelecidos mecanismos como a formação do pessoal ou o aprimoramento dos manuais, para que possíveis problemas de desempenho geral e departamental sejam resolvidos.

III. Recomenda-se realizar **validações regulares sobre a incidência do sistema Laura na qualidade dos dados clínicos** no registro digital do hospital.

IV. Em relação à explicabilidade algorítmica, sugere-se incorporar de modo sistemático e compreensível para um público geral as informações (Model Card - Mitchell et al., 2019) sobre o funcionamento do modelo algorítmico, incluindo: 1. Objetivos do sistema; 2. Dados; 3. Abordagem metodológica; 4. Descrição do algoritmo; e 5. Parâmetros de avaliação de desempenho e erros.

   A. Essas informações devem ser comunicadas, como parte da política de privacidade do hospital, a todos os pacientes cujos dados serão mensurados pelo sistema.

V. **Revisar o período de retenção dos dados pessoais** e seguir o princípio da minimização de dados, desde que sua eliminação não afete a qualidade do serviço nem a finalidade da tecnologia em questão.

   A. Realizar um treinamento em proteção de dados com os membros do sistema Laura e funcionários do hospital que aborde os requisitos básicos de proteção de dados sensíveis.

VI Examinar se a **segurança do sistema i**nclui o mecanismo de autenticação da identidade e incorpora chaves de acesso com informações conhecidas apenas pelo pessoal em questão. Integrar um mecanismo de registo e acompanhamento dos acessos ao sistema, com monitoramento de eventuais acessos não autorizados. Garantir a boa qualidade e proteção dos dados, por meio do file hashing e sistemas de proteção contra ataques. Garantir a criptografia adequada na comunicação e armazenamento dos dados.

VII. Revisar a **política de pseudonimização** de modo a assegurar o mais alto nível de confidencialidade e não reidentificação no âmbito da funcionalidade exigida.

No que diz respeito à **auditoria algorítmica focada na análise de dados**, este estudo começou analisando 2.874 registos hospitalares, que serviram de base para medir o impacto e o tratamento diferenciado por grupos. Assim, foram identificadas algumas características do conjunto de dados (*dataset*), como a predominância da população do sexo feminino, que corresponde a grupos mais jovens que a população do sexo masculino, ou a escassa presença de menores de 17 anos, que poderiam ajudar para explicar o comportamento do sistema. Com base nisso, de fato, tomou-se a decisão de eliminar esse grupo menor de registros da análise, pois poderia desviar os resultados.

Em seguida, foram realizadas três análises: uma comparação entre o risco observado e o previsto pelo sistema Laura para o mesmo grupo de pacientes, as taxas de FN e PPV para os mesmos grupos de idade e sexo, e a curva de calibração, a fim de estabelecer a relação entre valores preditivos e porcentagem de positividade por grupos.

Em síntese, **observam-se resultados consistentes quanto à menor capacidade preditiva do sistema Laura para pessoas do sexo feminino entre 18 e 39 anos.**

Tabela 14. Resumo dos resultados da análise de dados

Dimensão	Métricas	Resultado
Sexo	Risco	O sistema atribui um risco de morte previsto maior que o observado, com uma diferença de 14 pontos percentuais para ambos os sexos. Essa margem de risco dada pelo sistema pode decorrer de uma calibração que busca elevar o risco previsto, de forma a minimizar o risco de falsos negativos.
	PPV	O PPV é menor para o sexo feminino que para o masculino; ou seja, a probabilidade de a previsão de morte estar correta é menor para as mulheres que para os homens.
	FNR	São observadas baixas taxas de falsos negativos (~5-6%). O sistema Laura tende a raramente subestimar o risco de deterioração clínica. O FNRd do grupo do sexo masculino é de 0,83, já que o grupo feminino é o que apresenta a maior taxa de falsos negativos.
Idade	Risco	Assim como na dimensão sexo, o risco previsto é maior que o risco observado. Observa-se, também, que o risco previsto aumenta em cada faixa etária analisada. O grupo de 18 a 39 anos apresenta uma diferença de seis pontos percentuais (pp) entre o risco previsto e o observado; o grupo de 40 a 59 anos tem 10 pp; e o grupo de 60 anos ou mais tem 19 pp. O grupo com os menores riscos observado e previsto é a faixa etária de 18 a 39 anos.
	PPV	O PPV é maior para a faixa etária de 18 a 39 anos, o que pode prejudicar os grupos vulneráveis acima de 70 anos. A disparidade por grupos (PPVd) está dentro dos limites-alvo, embora a taxa de PPVd mais alta, que corresponde ao grupo de 18-39 anos, possa ser explicada por ser o grupo com o menor número de pacientes (360 vs. 1.550).
	FNR	São observadas baixas taxas de falsos negativos (~5-6%); o sistema Laura tende a raramente subestimar o risco de deterioração clínica. A disparidade por grupos mostra que a Laura tende a subestimar esse risco com mais frequência na faixa etária de 18 a 39 anos.

Sexo e idade	Risco	O risco previsto aumenta em pontos percentuais de modo semelhante ao observado na dimensão idade. Observa-se um maior aumento de pontos percentuais para o sexo feminino, ou seja, a previsão para o sexo masculino está mais próxima dos dados observados que para o sexo feminino.
	PPV	São observados valores mais elevados de PPVd nos grupos do sexo masculino em relação ao sexo feminino; a maior discrepância encontra-se entre os grupos de 18-39 anos do sexo masculino (1,35) e 18-39 anos do sexo feminino (0,78), estando esses valores fora dos limites-alvo (0,8 - 1,25). Essa diferença pode ser explicada por um viés nos dados: o menor risco de
	FNR	São observadas faixas de taxas baixas de falsos negativos mais amplas (~5-8%), sendo que os dois valores mais altos correspondem a faixas etárias do sexo feminino. A disparidade por grupos ultrapassa o limite-alvo (1,25) para os grupos do sexo feminino entre 18-39 anos (1,43) e 40-59 anos (1,32). O sistema Laura tende a subestimar o risco de deterioração clínica com
-	Calibração	A curva de calibração está mais ajustada à calibração perfeita nos casos de pontuações mais baixas e mais altas, mas se perde próximo ao ponto médio, onde também se observam maiores diferenças entre homens e mulheres.

Fonte: elaboração própria.

RECOMENDAÇÕES DA ANÁLISE ALGORÍTMICA

I. Monitorar os grupos com poucos pacientes e eliminar os grupos com muito poucos pacientes, como os de menores de 17 anos, do conjunto de dados (dataset) analisado. Nessa linha, recomenda-se estudar os **casos de variáveis com prevalência** muito baixa na amostra, considerando-se que não podem ser modeláveis de forma robusta.

   A. Uma possibilidade nesse sentido é incorporar alertas quando o sistema os detecta.

II. Como o sistema tende a d**esproteger as pessoas do sexo feminino de 18 a 39 anos**, recomenda-se:

   A. **Alertar os administradores do sistema sobre essa característica.** Ou seja, alertar a equipe do hospital que o sistema subestima o risco para esse grupo.

III. Recomenda-se **buscar uma melhor calibração do modelo**, particularmente em torno do valor que é utilizado como ponto de corte. Isso também deve ser contrastado em relação a seu efeito nas taxas de PPV e FN nos grupos cruzados (idade e sexo) analisados neste documento.

IV. Garantir o **preparo necessário das trabalhadoras e trabalhadores** que interagem com o modelo durante o processo de alinhamento, incorporando informações sobre suas margens de precisão para os grupos auditados.

V. **Explicar** o objetivo do modelo aos trabalhadores hospitalares, bem como aos pacientes em geral, esclarecendo que não se trata de um sistema de decisão autônomo, mas apenas de reforço objetivo na tomada de decisão.

REFERÊNCIAS

Alshahrani, A., Jamal, A. e Tharkar, S. (2021). How private are the electronic health records? Family physicians' perspectives towards electronic health records privacy. *Journal of Health Informatics in Developing Countries*, *15*(1). Disponível em https://www.jhidc.org/index.php/jhidc/article/view/298

Article 29 Data Protection Working Party. (2014). Opinion 05/2014, on *Anonymisation Techniques*. Disponível em https://www.pdpjournals.com/docs/88197.pdf

Ash, J. S., Berg, M. e Coiera, E. (2004). Some unintended consequences of information technology in healthcare. *Journal of the American Medical Informatics Association, 11*(2): 104-112.

Baeza-Yates, R. (2018). Bias on the web. C*ommunications of ACM 61*(6): 54-61.

Bandeira da Silva, D., Schmidt, D., da Costa, CA, da Rosa Righi, R. e Eskofier, B. (2021). DeepSigns: A predictive model based on Deep Learning for the early detection of patient health deterioration. *Expert Systems with Applications*, *165*(5). Disponível em https://www.sciencedirect.com/science/article/abs/pii/S0957417420307004

Barocas, S. e Nissenbaum, H. (2014). Big Data's End Run around Anonymity and Consent. In: Lane, J., Stodden, V., Bender, S. e Nissenbaum, H. (Eds.), *Privacy, Big Data, and the Public Good: Frameworks for Engagement*, 44-75. Cambridge: Cambridge University Press. doi: 10.1017/CBO9781107590205.004

Barocas, S. e Hardt, M. (2017). *Fairness in Machine Learning*. Tutorial at NIPS. https://mrtz.org/nips17/

Barocas, S. e Selbst, A. (2016). Big Data's Disparate Impact, *California Law Review, 104*: 671-732. Disponível em http://www.californialawreview.org/wp-content/uploads/2016/06/2Barocas-Selbst.pdf

Batista, N. O. W., Coelho, M. C. R., Trugilho, S. M., Pinasco, G. C., Santos, E. F. S. e Ramos-Silva, V. (2015). Clinical-epidemiological profile of hospitalised patients in pediatric intensive care unit. *Journal of Human Growth and Development, 25*(2): 187-193. Disponível em https://dx.doi.org/10.7322/jhgd.103014

Bihorac, A., Ozrazgat-Baslanti, T., Ebadi, A., Motaei, A., Madkour, M., Pardalos, P. M., Lipori, G., Hogan, W. R., Efron, P. A., Moore, F., Moldawer, L. L., Wang, D. Z., Hobson, C. E., Rashidi, P., Li, X. e Momcilovic, P. (2019). MySurgeryRisk: Development and Validation of a Machine-learning Risk Algorithm for Major Complications and Death After Surgery. *Annals of Surgery*, *269*(4): 652-662.

Bins, R. (2018). Algorithmic Accountability and Public Reason. *Philosophy & Technology*, *31*(4): 543-556.

Bradman, K., Borland, M. e Pascoe, E. (2014). Predicting patient disposition in a pediatric emergency department. *Journal of Paediatrics and Child Health, 50*(10): 39-44. Disponível em https://dx.doi.org/10.1111/jpc.12011

Cardoso, L. T., Grion, C. M., Matsuo, T., Anami, E. H., Kauss, I. A., Seko, L. e Bonametti, A. M.

(2011). Impact of delayed admission to intensive care units on mortality of critically ill patients: A cohort study. *Critical Care, 15*(1): R28.

Caruana, R., Lou, Y., Gehrke, J., et al. (2015). Intelligible Models for healthcare: predicting pneumonia risk and hospital 30-day readmission. In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* Cham, Switzerland: Springer International Publishing AG, 1721-1730.

Castillo, C. (2018). Algorithmic Discrimination. Assessing the impact of machine intelligence on human behaviour: An interdisciplinary endeavour. *Proceedings of HUMAINT Workshop*. Disponível em https://arxiv.org/pdf/1806.03192.pdf

Chouldechova A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments, arXiv:1610.07524. Disponível em: https://arxiv.org/abs/1610.07524

Churpek, M. M., Yuen, T. C., Winslow, C., Meltzer, D. O., Kattan, M.W. e Edelson, D. P. (2016). Multicenter Comparison of Machine Learning Methods and Conventional Regression for Predicting Clinical Deterioration on the Wards. *Critical Care Medicine, 44*(2): 368-374.

Cilliers, L. (2017). Exploring information assurance to support electronic health record systems. *2017 IST-Africa Week Conference (IST-Africa),* Windhoek, Namíbia: 1-8. doi: 10.23919/ISTAFRICA.2017.8102363.

Cowgil, B. (2019). Bias and productivity in humans and machines. *Upjohn Institute Working Paper,* No. 19-309, W.E. Upjohn Institute for Employment Research: Kalamazoo. doi: 10.17848/wp19-309. Disponível em: https://research.upjohn.org/up_workingpapers/309/

Danks, D. e John London, A. (2017). Algorithmic bias in autonomous systems. *Proceedings of the 26th International Joint Conference on Artificial Intelligence.* AAAI Press: 4691-4697.

Di Martino, D., Masturzo, B., Paracchini, S., Bracco, B., Cavoretto, P., Prefumo, F., Germano, C., Morano, D., Girlando, F., Giorgione, V., Parpinel, G., Cariello, L., Fusè, F., Candiani, M., Todros, T., Rizzo, N. e Farina, A. (2019). Comparison of two "a priori" risk assessment algorithms for preeclampsia in Italy: A prospective multicenter study. *Archives of Gynecology and Obstetrics*, *299*(6): 1587-1596.

Duncan, B. B., Cousin, E., Naghavi, M. et al. (2020). The burden of diabetes and hyperglycemia in Brazil: A global burden of disease study 2017. *Population Health Metrics 18*(9). Disponível em https://doi.org/10.1186/s12963-020-00209-0

Dwork, C., Hardt, M., Pitassi, T., Reingold, O. e Zemel, R. (2012). Fairness through awareness. *Proceedings of the 3rd innovations in theoretical computer science conference*: 214-226.

Ferryman, K. e Pitcan, M. (2018). Fairness in precision medicine. *Data & Society*. Disponível em https://datasociety.net/library/fairness-in-precision-medicine/

Gillen, S., Jung, C., Kearns, M. e Roth, A. (2018). *Online learning with an unknown fairness metric*, arXiv: 1802.06936. Disponível em https://arxiv.org/abs/1802.06936

Goldstein, B. A, Navar, A. M., Pencina, M. J. e Ioannidis, J. P. A. (2017). Opportunities and challenges in developing risk prediction models with electronic health records data: A systematic review.

Journal of the American Medical Informatics Association, *24*(1): 198-208.

Gonçalves, L. S., Amaro, M. L. M., Romero, A. L. M., Schamne, F. K., Fressatto, J. L. e Bezerra, C. W. (2020). Implementation of an Artificial Intelligence Algorithm for Sepsis Detection. *Revista Brasileira de Enfermagem*. *73*(3): 1-5.

Green, M., Lander, H., Snyder, A., Hudson, P., Churpek, M. e Edelson, D. (2018). Comparison of the Between the Flags calling criteria to the MEWS, NEWS and the electronic Cardiac Arrest Risk Triage (eCART) score for the identification of deteriorating ward patients. *Resuscitation,* 123: 86-91.

Gulshan, V., Peng, L., Coram, M., et al. (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA, 316*(22): 2402-2410.

Haupt, C. E. (2019). Artificial Professional Advice. *Yale Journal of Law Technology. 21*(3): 55-77.

Heidari, H., Ferrari, C., Gummadi, K. e Krause, A. (2018). Fairness behind a veil of ignorance: A welfare analysis for automated decision making. In: Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N. e Garnett, R. (Eds.). *Advances in Neural Information Processing Systems 31*. Montreal, QC: Curran Associates, Inc.: 1265-1276.

Hoff, T. (2011). Deskilling and adaptation among primary care physicians using two work innovations. *Health Care Management Review, 36*(4): 338-348.

Holstein, K., Wortman Vaughan, J., Daumé III, H., Dudik, M. e Wallach, H. (2019). Improving fairness in machine learning systems: What do industry practitioners need? *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Artigo N.° 600: 1-16.

IBGE (2018). Sistema de Contas Regionais: Brasil 2018. *Contas Nacionais*, 77. Disponível em https://biblioteca.ibge.gov.br/visualizacao/livros/liv101765_informativo.pdf

Joynt Maddox, K. E., Reidhead, M., Qi, A. C. e Nerenz, D. R. (2019). Association of Stratification by Dual Enrollment Status With Financial Penalties in the Hospital Readmissions Reduction Program. *JAMA Internal Medicine*, *179*(6): 769-776.

Kalil, A. J. (2017). Avaliação do impacto na identificação de pacientes com risco de sepse após implantação de um robô cognitivo de gerenciamento de risco (ROBÔ LAURA®). Dissertação de Mestrado. Curitiba: Universidade Tecnológica Federal do Paraná.

Kalil, A.J., Dias, V. M. C. H., Rocha, C. C., Morales, H. M. P., Fressatto, J. L. e Faria, R. A. (2018). Sepsis risk assessment: A retrospective analysis after a cognitive risk management robot (Robot Laura) implementation in a clinical-surgical unit. *Research on Biomedical Engineering, 34*(4): 310-316.

Katurura, M. e Cilliers, L. (2016). The extent to which the POPI Act makes provision for patient privacy in mobile personal health record systems. In: *The conference proceedings of IST-Africa* 2016, 11-13 de maio. Durban: IST-Africa.

Kim, M. P., Reingold, O. e Rothblum, G. N. (2018). *Fairness through computationally-bounded awareness*. arXiv: 1803.03239. Disponível em https://arxiv.org/abs/1803.03239

Kobylarz Ribeiro, J. et al. (2020). A Machine Learning Early Warning System: Multicenter Validation in Brazilian Hospitals. *2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS).* Rochester, MN: 321-326.

Laudon, K. C. e Laudon, J. P. (2010). *Management Information Systems*. New Jersey: Pearson Education.

Lippert-Rasmussen, K. (2013). *Born Free and Equal? A Philosophical Inquiry into the Nature of Discrimination*. Oxford: Oxford University Press.

Loreto, M., Lisboa, T. e Moreira, V. P. (2020). Early prediction of ICU readmissions using classification algorithms. *Computers in Biology and Medicine, 118*(C).

Lubarsky, B. (2017). Re-identification of "Anonymized Data". *Georgetown Law Technology Review*, *2*(1): 202-213.

Lum, K. e Isaac, W. (2016). To predict and serve? *Significance,* 13, 14-19.

Maconachy, V. W., Schou, C. D., Ragsdale, D. e Welch, D. (2001). A model for Information Assurance: An Integrated Approach. In: *The proceedings of the 2001 IEEE Workshop on Information Assurance and Security*. United States Military Academy, West Point, NY, 5-6 de junho.

Madden, M. (2018). Need medical help? Sorry, not until you sign away your privacy. *MIT Technology Review,* 23 de outubro. Disponível em https://www.technologyreview.com/s/612282/need-medical-help-sorry-not-until-you-sign-away-your-privacy/

Miranda, J. O. F. et al. (2020). Factors associated with the clinical deterioration recognized by an early warning pediatric score. *Texto & Contexto Enfermagem*, 29: 1-12.

Mitchell, M. S., Wu, A., Zaldivar, P., Barnes, L., Vasserman, B., Hutchinson, E., Spitzer, I., Raji, D. e Gebru, T. (2019). Model Cards for Model Reporting. In: Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19). *Association for Computing Machinery,* New York, NY, 220–229. doi: https://doi.org/10.1145/3287560.3287596

Morgan, R., Lloyd-Williams, F., Wright, M. e Morgan-Warren, R. (1997). An early warning scoring system for detecting developing critical illness. *Clinical Intensive Care*, 8: 100.

Muralitharan, S., Nelson, W., Di, S., McGillion, M., Devereaux, P., Barr, N. e Petch, J. (2021). Machine Learning–Based Early Warning Systems for Clinical Deterioration: Systematic Scoping. *Review Journal of Medical Internet Research*, *23*(2).

Narayanan, A. (2018). Tutorial: 21 definitions of fairness and their politics [*Abstract and video*]. *Conference on Fairness, Accountability, and Transparency.* NYC, 23 de fevereiro.

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

Ohm, P. (2009). Broken promises of privacy: Responding to the surprising failure of anonymization. *UCLA Law Review*, 57, 1701-1777. Disponível em https://pages.uoregon.edu/koopman/courses_readings/phil407-net/ohm_broken_promises_privacy.pdf

Pimentel, M. A., Redfern, O. C., Malycha, J., Meredith, P., Prytherch, D. R., Briggs, J., Young, J. D., Clifton, D. A., Tarassenko, L. e Watkinson, P. J. (2021). Detecting deteriorating patients in hospital: Development and validation of a novel scoring system. *Americal Journal of Respiratory and Critical Care Medicine.* Disponível em https://www.atsjournals.org/doi/abs/10.1164/rccm.202007-2700OC

Price, W.N. (2017). Regulating Black-Box Medicine. *Michigan Law Review*, *116*(3): 421-474.

Ratwani, R. M., Fairbanks, J. R., Hettinger, A. Z. e Benda, N. C. (2015). Electronic health record usability: Analysis of the user-centered design processes of eleven electronic health record vendors. *Journal of the American Medical Informatics Association*, *22*(6): 1179-1182. https://doi.org/10.1093/jamia/ocv050

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1: 206-215.

Sendak, M., Elish, M. C., Gao, M., Futoma, J., Ratliff, W., Nichols, M., Bedoya, A., Balu, S. e O'Brien, C. (2020). The human body is a black box: Supporting clinical decision-making with deep learning. In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (*FAT* '20*). Association for Computing Machinery, Nova York, NY, 99-109. doi: https://doi.org/10.1145/3351095.3372827

Speicher, T., Heidari, H., Grgic-Hlaca, N., Gummadi, KP, Singla, A., Weller, A. e Bilal Zafar, M. (2018). A Unified Approach to Quantifying Algorithmic Unfairness. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, julho de 2018: 2239-2248. doi: 10.1145/3219819.3220046

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

Tsai, T. L., Fridsma, D. B. e Gatti, G. (2003). Computer decision support as a source of interpretation error. *Journal of the American Medical Informatics Association*, *10*(5): 478-483.

Tucker, K. M., Brewer, T. L., Baker, R. B., Demeritt, B. e Vossmeyer, M. T. (2009). Prospective evaluation of a pediatric inpatient early warning scoring system. *Journal for Specialists in Pediatric Nursing*, *14*(2): 79-85. Disponível em https://dx.doi.org/10.1111/j.1744-6155.2008.00178.x

Tume, L. (2007). The deterioration of children in ward areas in a specialist children's hospital. *Nursing in Critical Care*, *12*(1): 12-19. Disponível em https://dx.doi.org/10.1111/j.1478-5153.2006.00195.x

Turney, P. D. (1996). How to shift bias: Lessons from the Baldwin effect. *Evolutionary Computation, 4*(3): 271-295.

Ueno, R., Xu, L., Uegami, W., Matsui, H., Okui, J., Hayashi, H., et al. (2020). Value of laboratory results in addition to vital signs in a machine learning algorithm to predict in-hospital cardiac arrest: A single-center retrospective cohort study. *PLoS ONE*, *15*(7): 1-16.

Williams, B. (Ed.). (2017). *National Early Warning Score (NEWS) 2 – Standardising the assessment of acute illness severity in the NHS*.

Zfania, T. K., Yang, J.,. Rossetti, S C., Cato, K. D., Kang, M. J., Knaplund, C., Schnock, K. O., Garcia, J. P., Jia, H., Schwartz, J. M. e Zhou, L. (2020). Mining clinical phrases from nursing notes to discover risk factors of patient deterioration. *International Journal of Medical Informatics*, 135. Disponível em https://www.sciencedirect.com/science/article/abs/pii/S1386505619309682

eticas

BID
Melhorando vidas

REFERENCIAS 58

INTRODUCCIÓN

Este documento presenta el Informe Final de la auditoría algorítmica del sistema Laura, llevada a cabo por Eticas Research and Consulting[1]. Esta auditoría no solo aborda la justicia algorítmica en el modelo de procesamiento automatizado; también conlleva una evaluación de la deseabilidad, aceptabilidad y gestión de los datos en dicho sistema. El documento desarrolla los resultados de estos análisis sobre la base de una descripción de su modelo y un análisis de su marco social de implantación.

El sistema Laura es un **Robot Cognitivo/Gestor de Riesgos** que actúa en la **identificación temprana de los riesgos de deterioro clínico.** Activo desde 2016, el sistema Laura ha analizado más de 8,6 millones de visitas en 40 centros clínicos y hospitalarios de varios estados de Brasil. El sistema ha ido variando su modelo, desde un enfoque centrado en la identificación del riesgo de presentar sepsis en pacientes en situación de internación hospitalaria, a uno más integral, donde la evaluación se establece sobre el riesgo de desmejora clínica y deceso, a partir de parámetros similares. **Esta auditoría examina la aplicación Laura, en su versión 1.0**, creada en 2017.

El principal objetivo del sistema Laura es alertar tempranamente un deterioro clínico susceptible de deceso, con el efecto de reducir la mortalidad y los costos del servicio hospitalario a través del análisis predictivo[2]. Se trata de un sistema de Inteligencia Artificial que ofrece una clasificación del riesgo de empeoramiento clínico del paciente, tras analizar los indicadores de las últimas cinco recolecciones de constantes vitales del mismo. Este sistema de predicción de riesgo se ha contrastado con el **sistema de puntuación *Modified Early Warning Score***[3] (MEWS), usado como estándar para la detección temprana de deterioro clínico (Kobylarz et al., 2020). La plataforma inteligente se encuentra actualmente conectada en la nube a más de 40 hospitales brasileños que tienen diferentes historias clínicas electrónicas (EHR de sus siglas en inglés).

La **auditoría algorítmica de Laura** se ha enfocado en **explorar posibles riesgos de sesgos o discriminación algorítmica** en los resultados ofrecidos por el sistema y traducidos efectivamente en intervenciones clínicas por parte del personal hospitalario. En este sentido, cabe tener en cuenta que, entre los productos ofrecidos por Laura, que incluyen herramientas de Detección de Deterioro Clínico, Atención Primaria, Gestión de Protocolos y Perfil Epidemiológico, centraremos nuestra atención en las herramientas de detección de deterioro clínico y sus gestores de información.

Con este fin, en las **secciones 2 "Estado del arte" y 3 "Contexto social"** de este documento nuestro análisis sitúa al sistema Laura tanto en el marco de un estado del arte sobre los sistemas automatizados de predicción de deterioro clínico como en su contexto socioeconómico y cultural. Este análisis se realiza mediante un estudio de la literatura, estadísticas relevantes y entrevistas con desarrolladores del sistema, así como también con personal clínico y de enfermería a cargo de su implementación en hospitales[4]. Sobre esta base se establecen

1 El equipo de investigación se encontró conformado por Emma López y Mariano Martín Zamorano. El equipo contó con la asesoría del Dr. Carlos Castillo, de la Universitat Pompeo Fabra.

2 Véase presentación en https://www.laura-br.com/en/

3 Las puntuaciones de alerta temprana (MEWS, por sus siglas en inglés) se han desarrollado para mejorar los mecanismos de detección deterioro sobre la base de los parámetros fisiológicos en los pacientes de las salas del hospital (Morgan et al., 1997). La utilización de datos sobre el deterioro intrahospitalario y el paro cardíaco demostraron ir precedidos de un período de aumento de anomalías en los signos vitales (Williams, 2017). Más información sobre *EWS en https://en.wikipedia.org/wiki/Early_warning_score

4 Estas primeras entrevistas incluyeron: entrevista 1, con personal técnico (17-03-2021), y entrevista 2, con dirección y coordinación

hipótesis de sesgo algorítmico para ser medidas cuantitativamente en otra fase de la auditoría, mediante los resultados que ofrece el sistema a lo largo de un período específico.

En la **sección 4 "Estrategia metodológica"**, el documento describe la **metodología establecida** para la medición de discriminación y justicia algorítmica en el sistema. Esta metodología se ha planteado como una exploración enfocada en recabar evidencia indirecta sólida sobre sesgo algorítmico para las variables sexo biológico y edad. Este enfoque no agotará todas las vías de evaluación de este fenómeno, pero brindará instrumentos para su consideración y monitoreo más abarcador en el futuro. Dicha metodología está asimismo destinada al estudio de su impacto social en términos de usabilidad, deseabilidad y aceptabilidad, así como al análisis del tratamiento de datos personales.

Finalmente, el informe presenta los **resultados del análisis cualitativo y cuantitativo** del sistema y las recomendaciones relacionadas con el mismo. Dichos resultados son presentados en las **secciones 5 "Resultados del estudio operativo y de aceptabilidad", 6 "Resultados del análisis algorítmico", ¿Reflejan esto los párrafos de esta página? y 7 "Conclusiones y recomendaciones",** que abordan las cuatro dimensiones principales de la auditoría, la protección de datos personales, la aceptabilidad y usabilidad de Laura y la justicia algorítmica en la asignación de riesgo de deterioro clínico, para cerrar con unas conclusiones que incluyen recomendaciones derivadas específicas.

1. CÓMO FUNCIONA EL SISTEMA

PREDICCIÓN DE DETERIORO CLÍNICO

El Robot Laura es un sistema especializado para la **evaluación de deterioro clínico**. Consiste en una plataforma inteligente conectada en la nube a más de 40 hospitales brasileños. Estos hospitales cuentan con diferentes Historias Clínicas Electrónicas (EHR) (Kobylarz et al., 2020) que cada hospital almacena en sus propias bases de datos (entrevista online, 17-03-2021). Dichas EHR no están estandarizadas a nivel interhospitalario.

Mediante Inteligencia Artificial (IA) y aprendizaje automático, el sistema proporciona alertas tempranas al equipo de atención médica en forma de una tasa de riesgo y otras informaciones sobre la condición del paciente. Esta información, que refleja la condición clínica del paciente en

asistencial del sistema (28-03-2021).

tiempo real (Kobylarz et al., 2020), se monitorea a través de un panel en vivo donde se muestran las alertas médicas según se van produciendo (Figura 1). A través de esta comunicación (que también incluye pautas sobre cómo debe actuar el personal de atención al identificar cambios), el robot indica qué pacientes pueden tener un **riesgo alto, medio y bajo de deterioro clínico**.

Figura 1. Panel del sistema Laura

Fuente: Laura.

El funcionamiento del sistema Laura, paso por paso, lo resumen Kalil et al. (2018: 311) de la siguiente manera:

I. **Realiza minería de datos,** a distancia, sobre todas las bases de datos y equipos de generación de datos del hospital.

II. **Clasifica** registros anómalos, inconsistentes y defectuosos.

III. **Evalúa** dichos datos para la generación de alarmas de riesgo para cada paciente, con la intervención del médico especialista en algoritmos.

IV. **Organiza** estas alarmas según su frecuencia e importancia en áreas de riesgo. Esto lo traducen visualmente, al equipo de atención, paneles de gestión de vista instalados en la enfermería del hospital.

V. **Activa** de forma autónoma la comunicación funcional del espectro cuando la zona de riesgo más crítica está activa; los datos continúan advirtiendo sobre el daño. Esta función también gestiona el envío de SMS (Servicio de Mensajes Cortos) y correos electrónicos a los profesionales de la salud que están a cargo. De este modo, el sistema Laura llama la atención sobre el riesgo captado por el robot y anticipa la atención que debe dirigirse a los pacientes implicados.

Figura 2. Ciclo de funcionamiento de Laura

Fuente: Laura.

LAURA ASSISTANT

Laura cuenta con una **aplicación móvil** que actualmente es su foco de desarrollo, pues permite un acceso más frecuente y rápido a los datos por parte del personal médico y de enfermería.

Mediante dicha aplicación, el personal sanitario contará en el futuro con un **reporte continuo de las personas que están en riesgo**. El sistema desarrollado reporta diversos datos, como la cantidad de personas afectadas, su evolución a lo largo del tiempo y el tipo de intervenciones necesarias, en un modo más dinámico. También podrá informarse sobre la eficiencia de la intervención en relación con los pacientes de mayor riesgo (rojo) y con las ratios de atención de los pacientes, por períodos de tiempo. Además, desde el dispositivo móvil podrán **realizarse intervenciones** mediante la comunicación con el equipo y tomar decisiones sobre el paciente en cuestión (la propia interfaz ofrece opciones de intervención al personal médico). Asimismo, la aplicación ofrecerá un registro continuo de su seguimiento y centralizará todos los mensajes para el personal.

El suministro de dispositivos móviles debería fomentar un acceso más dinámico a los datos sobre los pacientes y una mayor interacción del personal médico. Se espera que esta información genere mayor conciencia sobre el rendimiento de los equipos y estimule el alcance de nuevas metas, al generar datos estadísticos estructurados.

Figura 3. Presentación de Laura Assistant

Fuente: Laura.

Como indicaron en una de nuestras entrevistas (18-03-2021), el equipo del sistema Laura considera que los aspectos procedimentales son cruciales para su efectiva utilización. En este sentido, el sistema se encuentra en un proceso de reconsideración de su metodología y ámbito de actuación, que ha venido transformándose desde su foco inicial en sepsis a su actual atención al conjunto del deterioro clínico, para servir como herramienta de soporte de la decisión clínica. De este modo, la empresa busca asegurar un seguimiento continuo y preciso del estado del paciente, que vaya más allá de los puntos de desmejora clínica. En este sentido, se quiere brindar ayuda y capacidad de administración al personal médico para el seguimiento de otros procesos de gestión y control hospitalarios.

CARACTERÍSTICAS Y BENEFICIOS

Actualmente, el sistema Laura se orienta a apoyar las actividades y el desempeño del personal hospitalario y a mejorar la infraestructura de los hospitales. En particular, el equipo de Laura ha determinado[5] estas características del sistema por grupo de interés:

- **Para la administración hospitalaria:**
    - reduce los costos generales de hospitalización;
    - mejora la eficiencia en la rotación de camas (más pacientes atendidos);
    - genera informes y muestra las tendencias en tiempo real;
    - favorece una transformación digital.

- **Para el equipo médico**:
    - predice el deterioro del paciente mediante el uso de inteligencia artificial;
    - realiza intervenciones más efectivas con decisiones basadas en datos;
    - genera informes y muestra tendencias en tiempo real con información clínica agrupada (cronología del paciente);
    - alerta y activa el equipo de respuesta rápida.

- **Para el equipo de enfermería:**
    - recibe la información;
    - produce advertencias tempranas para que el equipo de atención pueda actuar sobre los pacientes de mayor riesgo;

5 Véase https://empowering-changemakers.eu/wp-content/uploads/2020/02/Projeto-Laura-BR.pdf

- disminuye la sobrecarga de trabajo;
- reduce el estado de alerta y la fatiga relacionada con la sobrecarga de información;
- aumenta la eficiencia operativa del equipo;
- empodera al equipo de atención para priorizar sus tareas (información para la acción);
- genera informes y muestra las estadísticas en tiempo real y en línea.

- **Para los pacientes:**
  - perciben un control sistemático de su estado;
  - perciben un entorno hospitalario con tecnología moderna y avanzada.

MODELO ALGORÍTMICO

El sistema Laura de detección de deterioro clínico utiliza un **algoritmo de Potenciación de Gradiente o *Gradient Boosting***. Esa es una técnica de aprendizaje automático para problemas de regresión y clasificación usada para predecir eventos poco comunes (aquellos correspondientes a menos de 5 % del conjunto de datos o dataset). El sistema Laura tiene un modelo general de predicción que ha sido entrenado con 121.000 datos de visitas únicas hospitalarias entre 2016 y 2019, provenientes de seis hospitales en diferentes localizaciones geográficas de Brasil (Rio Grande do Sul, Paraná y Minas Gerais).

Los **datos de entrenamiento** usados son seis:

1. Signos vitales (temperatura, saturación de oxígeno, ratio de respiración, nivel de glucosa en sangre y presión arterial),
2. Sexo biológico,
3. Edad,
4. Sala,
5. Departamento donde está ingresado el paciente, y
6. Duración en días de estancia en el hospital.

El resultado por predecir es la **mortalidad hospitalaria**.

Para la creación de los datos de entrenamiento se considera una ventana temporal de 36 horas antes del resultado que va a predecirse. De esta ventana, se descartan las últimas 12 horas para prevenir sesgos en el modelo. Así pues, el modelo usa como datos de entrenamiento cinco colecciones de signos vitales en un periodo de 24 horas.

El modelo de aprendizaje automático se **adapta a las necesidades y condiciones de cada centro**. Esto ocurre en dos formas: reentrenando el modelo con datos locales, cuando están disponibles, y negociando la sensibilidad del modelo con el equipo médico local. Con este fin, el hospital o centro clínico donde se implementa el sistema Laura debe tener un protocolo predeterminado para la atención de pacientes en riesgo de deterioro clínico (incluidas las variables relacionadas con alteración de signos vitales), que permita entrenar el sistema en un entorno real (Gonçalves et al., 2020). Sobre esta base (que también incluye pautas acerca de

cómo debe actuar el personal de atención al identificar cambios), el robot indica qué pacientes pueden tener un riesgo alto, medio y bajo de deterioro clínico.

El equipo de **Laura reentrena un modelo específico para cada hospital** cuando este cuenta con datos históricos suficientes (cinco años) de sus pacientes. Si no es posible, se **usa el modelo general**. Independientemente de si se utiliza un modelo específico o el general, siempre hay un segundo proceso de calibración del modelo con el equipo médico local. El sistema introduce pruebas durante un periodo de tiempo y el equipo médico local decide el umbral a partir del cual una probabilidad se considera riesgo alto.

ANÁLISIS REALIZADOS SOBRE EL SISTEMA LAURA

El sistema se ha evaluado en diferentes ocasiones mediante el estudio de sus diferentes versiones. Los análisis se han focalizado en distintos modelos algorítmicos y la eficiencia de sus resultados. El trabajo de Kalil (2017) analizó retrospectivamente el impacto de la implantación del sistema Laura en el proceso de identificación y manejo de pacientes en riesgo de sepsis en una unidad clínico-quirúrgica. Comparó las ratios del tiempo medio de servicio (TMA), es decir, el tiempo medio de inserción de cualquier registro de datos en el sistema de historia clínica electrónica del paciente (evolución, datos vitales, prescripciones, pruebas de laboratorio), calculados en forma autónoma por el robot cognitivo, seis meses antes y seis meses después de su implantación en el sistema. El estudio no reveló cambios significativos en esta tasa, pero destacó el **potencial del sistema para la predicción de riesgo** mediante su capacidad de minería de datos.

Kalil et al. (2018: 312) confirmaron los resultados antes expuestos. El objetivo de este estudio fue evaluar el impacto de la implementación del sistema Laura en los procesos relacionados con la identificación y atención de pacientes con riesgo de sepsis en una unidad clínico-quirúrgica de un hospital privado de Curitiba-PR. Se examinaron los registros clínicos de 60 pacientes identificados con infección y/o sepsis en un período de seis meses antes y después de la implementación de dicha tecnología en el hospital y se evaluó el tiempo de asistencia promedio a partir de la lectura autónoma del robot. Las diferencias en el tiempo promedio/mediana hasta la prescripción de antibióticos desde el primer signo de infección identificado, con o sin sepsis, no fueron estadísticamente significativas (p = 0,85). En cuanto al tiempo de asistencia promedio, se observó una r**educción de 305 a 280 minutos** al comparar los períodos de seis meses antes y después de la implementación de la tecnología (p = 0.02).

La investigación de Gonçalves et al. (2020) abordó la implantación de Laura en los aspectos vinculados a la **interacción entre el personal de enfermería y tecnología** en un hospital filantrópico durante el año 2018. Mediante observación participante y entrevistas con actores clave, se analizó la administración y operatividad del sistema en su contexto de adopción en forma cualitativa. El sistema, todavía focalizado en la identificación de riesgo de sepsis, demostró que fue **adoptado en forma participativ**a por el personal de enfermería, potenciando y dinamizando la toma de decisiones en la identificación precoz de la sepsis. Como resultado de este trabajo se recomendó que todos los casos de alerta fueran analizados y validados por los profesionales de la salud del hospital.

El trabajo de Kobylarz et al. (2020) analizó 121.089 consultas médicas de seis hospitales diferentes y 7.540.389 puntos de datos. Los autores compararon los protocolos aplicados en las salas para la detección de deterioro clínico con seis métodos de aprendizaje automático escalables diferentes (tres modelos clásicos de aprendizaje automático, modelos basados en

ratios logísticos y probabilísticos, y tres modelos impulsados por gradientes, como LightGBM). Los resultados mostraron **una ventaja en AUC** (*Area Under the Receiver Operating Characteristic Curve*) de 25 puntos porcentuales en el mejor resultado del modelo de aprendizaje automático en comparación con los protocolos actuales. Al evaluar la hipótesis de la alternativa o el suplemento AI sistema de predicción del deterioro clínico en las salas de los hospitales, el estudio reveló que el algoritmo que **presenta mejores resultados es el LightGBM, con AUC de 0,961 y F1 de 0,671**. Este algoritmo muestra una mayor precisión que el sistema MEWS, con puntuaciones 0,697 AUC y 0,175 F1.

Como puede advertirse, estas investigaciones indican que el sistema Laura puede alcanzar una cierta precisión en la identificación de riesgo de sepsis o deterioro clínico y promover mejoras en los tiempos de administración de los registros hospitalarios de datos. Además, los estudios proporcionan datos que han permitido ajustes al modelo. Por otro lado, revelan el potencial del sistema para su implantación en entornos hospitalarios.

No obstante, estos estudios no han analizado el impacto diferencial del sistema sobre diferentes grupos sociales en función de las variables de precisión o efectividad utilizadas. La presente auditoría se centrará, de forma complementaria, en el estudio del impacto diferencial del sistema por grupos protegidos, utilizando diferentes metodologías para identificar la eficiencia de Laura.

2. ESTADO DEL ARTE

La utilización de sistemas basados en **técnicas de aprendizaje automático** está avanzando de modo acelerado en el sector sanitario (Sendak et al., 2020; Topol, 2019). Algunas de las implantaciones de estos sistemas han demostrado una elevada efectividad en la detección y pronóstico de enfermedades, por ejemplo, en el caso de la retinopatía diabética (Gulshan et al., 2016). No obstante, cabe considerar **posibles efectos no deseados de la automatización** de los análisis clínicos y diagnósticos, que incluyen la violación de la privacidad del paciente o la deshumanización en su tratamiento (Ferryman y Pitcan, 2018; Madden, 2018). Estas son algunas

de las razones por las que el uso de estos sistemas se encuentra estrictamente regulado por normativas nacionales e internacionales, que los hacen susceptibles de auditorías continuas (Haupt, 2019; Price, 2017).

Este apartado analizará la literatura que aborda sistemas similares a Laura y sus implicaciones, con el fin de considerar el alcance general y las principales limitaciones de dichos sistemas.

AUTOMATIZACIÓN DE RIESGO DE DETERIORO CLÍNICO

Existe una serie de factores **estructurales, tanto tecnológicos como vinculados a los recursos humanos**, que incide en las tasas de mortalidad hospitalaria en forma significativa. Los pacientes gravemente enfermos suelen tener cambios en sus signos vitales durante un período de tiempo antes de empeorar. La falta de capacidades técnicas y humanas en la detección temprana de aquellos pacientes que requieren un tratamiento prioritario ha demostrado tener efectos negativos en este proceso diagnóstico, derivando en muchos casos en un aumento de las tasas de decesos y desmejora clínica (Pimentel et al., 2021; Goldstein et al., 2017).

En las últimas décadas se han desarrollado **diferentes sistemas automatizados** para la identificación y notificación de los **primeros signos de deterioro clínico** y fisiológico (Goldstein et al., 2017). La adopción de historias clínicas electrónicas mejoró la disponibilidad de datos, que pueden procesarse mediante técnicas de aprendizaje automático para extraer información que respalde las decisiones clínicas. Los modelos de aprendizaje automático más utilizados con este fin incluyen la regresión logística, métodos basados en árboles, métodos basados en kernel y redes neuronales (Muralitharan et al., 2021). Un modelo algorítmico basado en *Track and Trigger Scoring System (TT)* y que es clave en el reconocimiento temprano, es el ***Modified Early Warning Score*** **(MEWS)**. Se ha demostrado que su implantación en ciertos contextos perfecciona los mecanismos hospitalarios destinados a monitorear los signos vitales de los pacientes. Asimismo, se ha sugerido que sirve como soporte para el personal de enfermería en la identificación de pacientes en situación de riesgo, pues ayuda a asegurar su situación clínica.

Otros modelos, desarrollados mediante la técnica *deep learning* y que emplean *Recurrent Neural Networks*, the *Long Short-Term Memory*, se han usado con éxito para predecir los signos vitales del paciente y la posterior evaluación de la gravedad de su estado de salud, utilizando Índices de Pronóstico (con una precisión de 80 %) (Bandeira da Silva et al., 2021). Valiéndose de este modelo es posible predecir futuros diagnósticos graves que no serían identificados mediante el análisis de los signos vitales del paciente en su situación presente. Otros sistemas de *deep learning* (redes neuronales) se usan para la detección de pacientes con riesgo de paro cardíaco, demostrando así una alta sensibilidad y una baja tasa de falsas alarmas (Ueno et al., 2020). Las distintas finalidades clínicas de estos sistemas en el ámbito hospitalario y el enfoque de enfermería en su aplicación se están estudiando activamente. En muchos casos muestran que contribuyen a un mejor monitoreo de los signos vitales del paciente y su seguridad clínica.
Los **datos de entrada** de estos sistemas son múltiples y dependen, entre otros factores, de la definición empleada de deterioro clínico. Algunos han demostrado cierta eficiencia al identificar riesgos mediante el análisis del lenguaje natural, utilizando las notas de enfermeras y enfermeros incluidas en los registros hospitalarios (*electronic health record*s (EHR) (Zfania et al., 2020). Las predicciones basadas únicamente en los atributos recopilados en la admisión hospitalaria han demostrado ser muy precisas para predecir el riesgo de readmisión en las Unidades de Cuidados Intensivos (Loreto et al., 2020), cuyo estudio sugiere que los "marcadores tempranos" pueden ser particularmente útiles para la predicción de riesgo de deterioro clínico en pacientes con alto riesgo de deterioro clínico después del alta de la UCI.

No obstante, los diferentes algoritmos que se están utilizando para detectar deterioro clínico han mostrado **diferentes grados de eficiencia**, dependiendo del contexto hospitalario y social, así como los predictores de riesgo utilizados. Por ejemplo, en función de cada contexto social, terapéutico y organizacional de implantación, el algoritmo de random forest o los algoritmos de regresión logística han sido más precisos en la identificación de riesgo de deterioro clínico (Churpek et al., 2016). Además, estos autores demuestran la importancia de una buena calibración y del "*gradient boosting*" en predictores similares.

Otras posibles fuentes de sesgos en estos sistemas se relacionan con el **diseño del modelo predictivo**. Un sistema de detección de pacientes en deterioro clínico en internación que utiliza un esquema de puntuación se modificó durante su desarrollo y validación para reducir riesgos de sesgos **contra los pacientes mayores** (Pimentel et al., 2021). Mientras en ciertas condiciones los pacientes mayores de 80 años tienen una probabilidad decreciente de sufrir un paro cardíaco o de ser transferidos a la UCI, los resultados de esta investigación mostraron una variación más amplia en el riesgo global previsto para los pacientes mayores de 80 años. La solución planteada a este problema fue incluir "una amplia gama de factores del paciente (comorbilidades, fragilidad)" en el modelo (Pimentel et al., 2021: 18).

Un estudio reciente, que analiza los resultados brindados por diversos predictores de riesgo en el contexto hospitalario, evidencia la necesidad de considerar los **posibles sesgos integrados en los datos de los EHR** como, por ejemplo, ausencia o calidad de datos para ciertas variables (Goldstein et al., 2017). En este sentido, cabe tener en cuenta que la codificación electrónica de ciertos datos (por ejemplo, con respecto a decisiones clínicas) varía entre hospitales y, en muchos casos, no es lo suficientemente sólida para su inclusión en un modelo generalizable (Pimentel et al., 2021).

Finalmente, la utilización de estos sistemas puede verse afectada por la **percepción de su utilidad y sesgos humano**s integrados en el uso de los sistemas o en los protocolos de incorporación de datos. Por ejemplo, la literatura ha evidenciado que la admisión a la UCI varía en función de la experiencia de los médicos y su percepción de los beneficios y los factores organizativos (por ejemplo, la disponibilidad de camas) (Green et al., 2018).

De este modo, el **modelo algorítmico y su base teórica**, los **sesgos históricos presentes en los datos** de entrada, los **procesos de aprendizaje** automático y el **sesgo en el uso** son las principales fuentes de discriminación algorítmica que hay que considerar también en el estudio de Laura.

3. CONTEXTO SOCIAL

Esta sección describe sucintamente el **contexto social de implantación de Laura**, actualmente utilizado en centros clínicos y hospitalarios de diferentes estados del sur de Brasil. Si bien esta auditoría no se orienta a contrastar en forma empírica y comparativa posibles sesgos del sistema Laura en una muestra extensa de hospitales a lo largo de Brasil, este apartado persigue establecer un marco general para su análisis e ilustrar diferencias sociales estructurales que pueden integrarse al sistema en forma de discriminaciones no deseadas.

La República Federativa de Brasil la componen **26 estados y un Distrito Federa**l, donde se encuentra Brasilia, su capital. Los estados están organizados en cinco regiones geográficas (Norte, Noreste, Sureste, Sur y Centro-Oeste), que tienen importantes diferencias económicas, culturales y demográficas. El país cuenta con un estimado de **209 millones de habitantes** (al año 2018).[6] Mientras la expectativa de vida ha aumentado desde el Censo de 1999, las tasas de natalidad vienen disminuyendo desde hace décadas y han caído por debajo de dos hijos por mujer.[7] De este modo, se advierte un paulatino envejecimiento poblacional, aunque continúa siendo un país de amplia población por debajo de 50 años. Según el Censo del año 2010, la población indígena brasileña es de 896.917 habitantes, lo que equivale a 0,5 % de la población total del país[8].

La pobreza y la desigualdad social son significativas en el país. La población pobre se acerca a 20 % de la población si se considera la línea de pobreza para la clase de ingresos medios-altos (13,8 en reales brasileños en 2018) por día/persona.[9] En 2010, los estados de las regiones Sureste, Sur y Centro-Oeste tenían índices de desarrollo humano (IDH) altos o muy altos (por encima de 0,699), mientras que el **Noreste y el Norte tenían índices de nivel medio (0,600 y 0,699, respectivamente)**. Según datos de las Naciones Unidas, el país en su conjunto presenta un IDH alto (entre 0,700 y 0,7999)[10]. No obstante, mientras que la mayoría de los estados del sur tienen un PIB per cápita por encima de 10.000 dólares alcanzan 13.299 dólares en el caso de Sao Paulo, muchos de los estados del norte se encuentran por debajo de 7000 dólares (IBGE, 2018).

El **Sistema Único de Salud de Brasil** (SUS, por su sigla en portugués) se encarga de las políticas orientadas a garantizar el acceso universal e integral a los servicios de salud. Algunos de sus objetivos son la promoción de la equidad, la gestión descentralizada y la participación social. La gestión del sistema la comparten los tres niveles de gobierno: el Ministerio de Salud en el nivel federal y las secretarías de salud estatales y municipales en los niveles inferiores. El sistema se financia con impuestos y contribuciones en los niveles federal, estatal y municipal[11]. No obstante, el SUS tiene una cobertura limitada o territorialmente desigual, lo que puede incidir en las tasas de deterioro clínico y el seguimiento médico de los pacientes. Por ejemplo, se ha destacado que la disponibilidad de camas libres en UCI es un problema muy importante y generalizado en el país (Cardoso et al., 2011).

Las variables sociodemográficas mencionadas (pobreza, índices de natalidad) y el alcance del sistema sanitario nacional son algunos de los factores que pueden condicionar estructuralmente el riesgo de deterioro clínico de las personas hospitalizadas en las distintas ciudades y pueblos del país. Dadas las características del modelo de Laura, mayormente basado en datos clínicos (temperatura, saturación de oxígeno, ratio de respiración, nivel de glucosa en sangre y presión arterial), pero también demográficos y hospitalarios (sexo biológico, edad, sala, departamento donde está ingresado el paciente y duración en días de la estancia en el hospital), cabe preguntarse por los posibles sesgos **derivados de su implementación en organizaciones y contextos sociales específicos**. La compleja relación entre estos factores y las posibles

6 Véase: World Bank data. https://datatopics.worldbank.org/world-development-indicators/

7 Véase: Pan American Health Organization, based on data from the United Nations Department of Economic and Social Affairs Population Division. New York; 2015.

8 Véase: Brazilian Institute of Geography and Statistics – IBGE. Indígenas. Disponible en: http://indigenas.ibge.gov.br/graficos-e-tabelas-2.html

9 Véase: https://databank.worldbank.org/data/download/poverty/33EF03BB-9722-4AE2-ABC7-AA2972D68AFE/Global_POVEQ_BRA.pdf

10 A nivel municipal, casi 80 % de la población vivía en municipios con IDH bajo o muy bajo en 1991; en 2010, sin embargo, esa proporción se había reducido a 11 %. Véase: United Nations Development Program – UNDP. Atlas. Disponible en http://www.atlasbrasil.org.br/2013/pt/o_atlas/idhm/

11 Según el Instituto Brasileño de Geografía y Estadística, el gasto total en salud en 2013 ascendió a 8 % del PIB del país, con un 3,6 % de gasto público.

diferencias en los niveles de riesgo de deterioro clínico esperables para distintos grupos sociales puede ilustrarse, por ejemplo, en la distribución territorial y el ritmo de crecimiento de la prevalencia y mortalidad por diabetes, que son más elevados en las regiones Norte, Noreste y Centro-Oeste del país (Duncan et al., 2020). Existen también otras variables grupales o intersectadas que pueden actuar como predictores de deterioro clínico. Por ejemplo, un estudio con 271 niños y niñas realizado en el Hospital Estadual da Criança da Bahia reveló que el sexo masculino ha sido más prevalente entre los menores que presentan deterioro clínico (Miranda et al., 2020), lo que confirma estudios que demuestran que este grupo prevalece en los ingresos en UCI y con respecto a afecciones respiratorias (Batista et al, 2015; Time, 2007).

Teniendo en cuenta lo anterior, cabe considerar que **existen predictores de riesgo de deterioro clínico que pueden favorecer la discriminación algorítmica debido a sesgos históricos** (datos reales que se ven afectados o arraigados en cuestiones de discriminación, legado o políticas injustas). Dado que el sistema Laura lo han implantado diferentes instituciones públicas y privadas del sur del país, conviene tener en cuenta estos elementos en el análisis de aquellos sesgos que derivan en un impacto diferencial sobre ciertos grupos poblacionales, cuando pueden trascender la capacidad de modelización de riesgo establecida en cada centro en función de sus condiciones sociodemográficas y clínicas específicas.

4. ESTRATEGIA METODOLÓGICA

JUSTIFICACIÓN TEÓRICO-METODOLÓGICA

Uno de los ejes fundamentales de la auditoría algorítmica es la **identificación y el análisis de la discriminación algorítmica**. Esta sección delimitará este concepto, proporcionando las definiciones clave para identificar y analizar las diferentes formas de sesgo algorítmico injusto o discriminatorio.

Para enmarcar el sesgo algorítmico, primero debe distinguirse entre diferentes formas de

discriminación. Siguiendo las definiciones de Lippert-Rasmussen (2013), la discriminación genérica ocurre cuando alguien trata a una persona A peor de lo que trataría a otra persona B, porque A tiene algún atributo que B no tiene. La **discriminación grupal ocurre cuando dicho atributo consiste simplemente en pertenecer a un grupo socialmente destacado**, es decir, un grupo en el que la membresía "es importante para la estructura de las interacciones sociales en una amplia gama de contextos sociales" (Lippert-Rasmussen, 2013: 30). Requiere, asimismo, animosidad contra un grupo la creencia de que las personas pertenecientes a este grupo son inferiores o la creencia de que dichas personas no deberían mezclarse con las demás.

En esta línea, **para ser considerado discriminatorio, el sesgo debe involucrar a uno o más de los llamados grupos protegidos,** que corresponden fundamentalmente a los atributos protegidos resumidos en la Tabla 1. Esta síntesis se basa en los atributos protegidos. contemplados en la Ley de Igualdad del Reino Unido 2010[12] (Sección 4), y en la Carta Europea de los Derechos Fundamentales[13]. Cabe señalar que esta lista no es exhaustiva, porque puede adaptarse o modificarse, según el contexto[14]:

Tabla 1. Grupos y atributos (legalmente) protegidos

Grupos protegidos (no exhaustivo)	Atributos protegidos
Niños y ancianos	Edad
Personas discapacitadas (físicas y mentales)	Discapacidad
Mujeres y transexuales	Género o reasignación de género
Embarazadas	Embarazo
Musulmanes, judíos	Religión o creencia
Gais, lesbianas, bisexuales, intersexuales...	Orientación sexual
Personas con bajos ingresos/escasos recursos	Propiedad/Recursos materiales

Fuente: elaboración propia.

La discriminación estadística es una **discriminación grupal basada en un hecho que es estadísticamente relevante**. Un ejemplo clásico de discriminación estadística es no contratar a una mujer que reúne las competencias para un puesto laboral porque las mujeres tienen mayor probabilidad de tomar una licencia de maternidad. En cambio, la discriminación no estadística ocurre cuando la mujer no es contratada porque ha dicho que tiene la intención de tener un hijo y, en consecuencia, tomar una licencia de maternidad (Lippert-Rasmussen, 2013). Si se ignora la animosidad correspondiente a los seres humanos, pero no a los algoritmos, y se considera que cualquier característica utilizada en el aprendizaje de máquina como estadísticamente relevante, podrá decirse que los algoritmos pueden discriminar (Castillo, 2018).

12 Véase información detallada en https://www.gov.uk/guidance/equality-act-2010-guidance

13 Legislación disponible en https://eur-lex.europa.eu/legal-content/ES/TXT/?uri=LEGISSUM%3Al33501

14 Los grupos desfavorecidos pueden definirse en relación con los atributos mencionados en el artículo 21 (No discriminación), de la Carta Europea de los Derechos Fundamentales: "sexo (y género), raza, color, origen étnico o social, características genéticas, idioma, religión o creencia, opinión política o de cualquier otro tipo, pertenencia a una minoría nacional, propiedad, nacimiento, discapacidad, edad u orientación sexual". Estos grupos protegidos se definen como individuos y grupos que comparten una o más de las 'características protegidas'.

Debe tenerse en cuenta que las definiciones antes brindadas se diferencian de las definiciones estándar de sesgo estadístico, que implican distorsiones de un cálculo estadístico resultante de muestras sesgadas o estimativos cuyo cálculo es incorrecto en relación con el valor correcto o esperado de un parámetro (Turney, 1996). Por lo tanto, el sesgo estadístico no puede (siempre) ser un criterio adecuado de equidad algorítmica. Aunque al menos normativamente, el sesgo y la discriminación pueden ser justos o injustos, esto dependerá de cómo se interpreten los resultados social y éticamente. Siguiendo la lógica anterior, una definición **más precisa de sesgo algorítmico —o discriminación algorítmica— implica la producción sistemática de resultados desventajosos contra grupos socialmente destacados, particularmente grupos desfavorecidos**. Este sesgo está incrustado en las propiedades matemáticas de un algoritmo.

El sesgo algorítmico se ha dividido en dos tipos diferentes, según la etapa del proceso de aprendizaje automático en el que sucede (Danks y London, 2017). En primer lugar, el sesgo algorítmico, así como los modelos sesgados, pueden estar **sesgados debido a la recopilación y el uso de datos de entrenamiento sesgados** al entrenar o modelar algoritmos durante las etapas iniciales de desarrollo (Cowgill, 2019). En segundo lugar, el **sesgo posalgorítmico o de procesamiento** se relaciona con el modelado del sistema causado por sus interacciones con los usuarios —procesamiento posterior en el gráfico que aparece a continuación—.

En este caso, el llamado tratamiento dispar de los subgrupos puede basarse en una lógica aparentemente razonable, pero que de todos modos conduce a la discriminación (Barocas y Selbst, 2016). Por lo tanto, la interpretación del usuario del resultado del procesamiento algorítmico y el contexto social son claves para evaluar si es justo o injusto (Baeza-Yates, 2018). Las diferentes fases durante las cuales puede ocurrir el sesgo algorítmico, que son las mismas fases en las que puede mitigarse el sesgo algorítmico, se resumen en la siguiente imagen.

Figura 4. Etapas en que puede mitigarse el sesgo algorítmico

Fuente: Hajian, S., Bonchi, F., y Castillo, C. (2016).

LA JUSTICIA ALGORÍTMICA

La definición y la sistematización de la equidad algorítmica se han convertido en temas vitales para desarrolladores y académicos en este campo (Gillen et al., 2018). En términos generales, la falta de equidad algorítmica podría definirse como cualquier caso "donde los sistemas AI/ML funcionan de manera distinta para diferentes grupos de maneras que pueden considerarse indeseables" (Holstein et al., 2019: 3). A pesar de que se han desarrollado **métodos cuantitativos para captar y medir el tratamiento/impacto dispar sobre los grupos desfavorecidos**, estas técnicas no pueden abarcar el debate sobre qué grupos pueden considerarse desfavorecidos y

qué puede considerarse tratamiento diferencial en un determinado contexto sociocultural. De hecho, la literatura ha demostrado la incompatibilidad habitual entre los modelos estadísticos de equidad y las interpretaciones hechas por usuarios o ciudadanos (Binns, 2018; Kyung Lee, 2018). Este debate se manifiesta **en múltiples definiciones de equidad**, lo que hace difícil alcanzar una única definición aceptada para ser utilizada por científicos e ingenieros. Narayanan (2018), por ejemplo, ha identificado 21 definiciones de justicia algorítmica.

Una de las definiciones más importantes se refiere a la **equidad grupal, que implica que el sistema algorítmico vigente no debería tratar de manera injusta a grupos sociales específicos**. Entre las medidas de equidad grupal se destacan las tres básicas descritas por Barocas y Hardt (2017): **independencia** (también conocida como paridad demográfica o paridad estadística), **separación** (conocida como probabilidades igualadas o evitación del maltrato desigual) y **suficiencia** (o calibración), que son tres de las más utilizadas en la literatura.

- **Independencia** significa que la probabilidad de asignar un resultado es independiente del atributo protegido (por ejemplo, en la predicción de reincidencia, cuando la raza es el atributo protegido; implica que la fracción de individuos asignados por un algoritmo a la clase de alto riesgo será la misma, independientemente de la raza).

- **Separación** significa que la probabilidad de asignar un resultado es independiente del atributo protegido, dado el resultado real (por ejemplo, en la predicción de reincidencia, la fracción de individuos asignados por un algoritmo a la clase de alto riesgo será la misma en todas las razas entre los individuos que no cometan un nuevo delito en el futuro).

- **Suficiencia** significa que el resultado asignado por un algoritmo no necesita combinarse con atributos protegidos para obtener una predicción (por ejemplo, en la predicción de reincidencia, que un puntaje dado se traduzca en la misma probabilidad de cometer un delito, independientemente de la raza).

Algunas de estas métricas de equidad grupal pueden ser incompatibles entre sí. Por ejemplo, al analizar sistemas para predecir la reincidencia, Chouldechova (2017) reveló que un instrumento que satisfaga la paridad predictiva no puede tener las mismas tasas de falsos positivos y negativos entre todos los grupos cuando la prevalencia de reincidencia difiere entre dichos grupos.

Además, como indican Heidari et al. (2018), las nociones estadísticas de equidad no garantizan la equidad a nivel individual. De hecho, una noción diferente de equidad algorítmica grupal es la de equidad algorítmica individual, que establecieron por primera vez Dwork et al. (2012) y que habla sobre un trato consistente a los individuos.

Para que un sistema sea justo desde un punto de vista individual, **dos individuos —que son similares en términos de los objetivos y el modelo del algoritmo— deben recibir resultados similares**. Esto ocurre también si son similares con respecto a sus características en la realidad, porque puede que el modelo considere iguales a dos individuos que en la realidad no lo son, al utilizar variables irrelevantes o incorrectas sobre ellos. Por lo tanto, este modelo impone restricciones en el tratamiento para cada par de individuos (Kim et al., 2018). Sin embargo, como señalan Speicher et al. (2018), estas métricas no tienen en cuenta factores contextuales más amplios, como las diferencias en actividades anteriores realizadas por cada individuo o el poder económico o social de cada uno de ellos. Además, según Speicher et al. (2018), no existen mecanismos computacionales eficientes para integrar este tipo de enfoques conceptuales. También, un sistema puede cumplir el criterio de equidad individual, pero generar un resultado

consistentemente adverso para un grupo determinado de individuos.

De este modo, existe un debate en curso en la literatura académica sobre el desarrollo de métricas de equidad adaptadas a diferentes tipos de algoritmos y sistemas. Además, se observa una **relación compleja entre equidad y eficiencia**, pues en algunos casos la precisión predictiva puede verse perjudicada con el objetivo de mejorar un sistema en términos de equidad (Narayanan, 2018). De hecho, cualquier enfoque metodológico adoptado con el propósito de evaluar el sesgo debe combinar el análisis de factores específicos que determinan la equidad, de manera que permitan hacer una contextualización sociológica y lograr los objetivos del procesamiento algorítmico. Para esto, el contexto social en el que opera el sistema debe tenerse en cuenta, tanto desde el punto de vista cuantitativo como cualitativo.

En este sentido, este informe sigue un esquema propuesto por Castillo (2018), según el cual los métodos algorítmicos que utilicen algún criterio para ordenar elementos como personas, grupos o similares, deberían poder alcanzar la equidad en términos de estos factores:

1. Una presencia suficiente de elementos del grupo protegido.
    a. Ausencia de discriminación estadística (grupal)
    b. Prevención de daños asignativos a un grupo
2. Un tratamiento consistente de los elementos de ambos grupos.
    a. Ausencia de discriminación individual.
3. Una representación adecuada de los grupos desfavorecidos.
    a. Prevención de daños de representación en un grupo

ENFOQUE METODOLÓGICO PARA LAURA

Teniendo en cuenta las definiciones antes descritas, esta sección describe brevemente las fases de la auditoría algorítmica de Laura y la metodología utilizada para evaluar la discriminación algorítmica.

FASES DE LA AUDITORÍA ALGORÍTMICA

La auditoría se compone de cuatro fases:

1. Estudio preliminar
Esta primera fase se dedicó principalmente a recabar información básica sobre las partes implicadas en el diseño, el desarrollo y la implementación del modelo, el modelo en sí mismo y su integración en las dinámicas propias de las organizaciones que representan las partes involucradas.

Para ello, se estableció contacto con las personas responsables del sistema Laura. Estas actividades permitieron recabar información básica sobre el sistema y las necesidades detectadas para su desarrollo e implementación. Toda la información necesaria para la realización de la auditoría, así como las decisiones tomadas por el equipo auditor de aquí en adelante, se reflejan en este y otros documentos de trabajo interno.

2. Mapeo de la situación

En segundo lugar, el equipo auditor estableció cómo, cuándo, por qué y para qué desarrolló e implementó este algoritmo en concreto. Asimismo, mediante una *Model card* con datos sobre el sistema se examinó si este cumplía un listado de requisitos básicos para poder ser auditado y si las partes responsables de su diseño, desarrollo e implementación tenían la disposición de proporcionar la información necesaria para su realización.

3. Plan de análisis
Esta fase consiste en definir y consensuar con el cliente los términos (cómo y para qué) y los plazos estimados (cuándo) para el desarrollo de la auditoría.

Con este objetivo, se celebraron varias reuniones e intercambios de información con las personas responsables de Eticas Research and Consulting y de Laura. Sobre esta base, el equipo auditor definió un plan de análisis (03-2021), para poner en común y consensuara las partes implicadas. Con base en lo acordado, se elaboró y entregó una propuesta del plan de análisis y se definió el equipo auditor, con conocimiento específico sobre el sistema en cuestión.

4. Análisis e informe final
Esta fase se centró en la ejecución del Plan de análisis, dentro de un cierto margen de flexibilidad respecto a lo planeado, en función de las circunstancias del estudio. En este caso, el análisis correspondió con lo estipulado a continuación.

En términos generales, la metodología de auditoría algorítmica de Eticas R&C se realiza en dos partes complementarias, cuyo objetivo es comprender la complejidad del modelo y sus posibles implicaciones:

- Por un lado, un **estudio de carácter cualitativo**, con el objetivo de comprender las implicaciones del sistema y su implementación, en el contexto socioeconómico, técnico y organizacional en el que se inscribe.
- Por otro, un **estudio de tipo cuantitativo**, basado en técnicas de análisis estadístico y ciencia de datos, principalmente enfocado en detectar y recomendar medidas de corrección para posibles casos de imprecisión, discriminación, tratamiento o impacto diferencial o sesgo algorítmico, provocadas por el sistema. Cabe señalar que en el caso de Laura se ha seguido una política de protección de datos que ha hecho necesario trabajar con registros completamente anonimizados, también en lo que respecta a la confidencialidad del hospital analizado.

En esta fase de la auditoría se realizan los análisis planificados para el entregable final y se extraen sus resultados principales.

APLICACIÓN DEL PLAN DE ANÁLISIS ALGORÍTMICO

El análisis del sistema automatizado y el estudio de su sesgo e impacto diferencial por grupos se realizó siguiendo una metodología que va del mapeo del sistema algorítmico y sus datos de entrada/entrenamiento hasta la aplicación de métricas orientadas a establecer diferencias estadísticas en línea con los criterios de suficiencia e independencia mencionados en el apartado anterior.

Hay cuatro pasos principales en la detección del sesgo algorítmico:

(1) definir la asignación de elementos a grupos;

(2) definir los grupos protegidos;

(3) determinar un conjunto de métricas destinadas a medir el sesgo, y

(4) medir y comparar entre grupos.

El primer paso simplemente **clasifica los elementos de datos en grupos**, que pueden estar superpuestos (asignación *soft*) o no superpuestos (asignación *hard*). Dicha superposición se refiere a la convergencia de más de una característica protegida que ha de considerarse; por ejemplo, mujer con bajos ingresos. En la mayoría de los casos, los datos reflejarán datos correspondientes a cada una de las personas y, por lo tanto, los grupos se realizarán según características individuales. Puede utilizarse cualquier característica asignada a múltiples individuos para crear tales grupos, pero se presta especial atención a las características protegidas antes mencionadas. Estas agrupaciones se crean en los datos para evaluar en qué medida un algoritmo puede tratar o afectar a un grupo de manera diferente a otro.

El segundo paso determina **cuáles grupos se han definido como protegidos**, lo que significa que no deben verse desfavorecidos por la aplicación del algoritmo y que el impacto de los algoritmos en ellos será monitoreado de manera especial. En algunos casos, los grupos protegidos pertenecen a categorías que están legalmente amparadas (por ejemplo, personas con discapacidades). En otros casos, la definición de lo que constituye un grupo protegido se relaciona con un compromiso que puede no ser legalmente vinculante, como la intención de aumentar la participación de mujeres o minorías que podrían estar subrepresentadas en ciertos puestos. Una definición adicional de grupo protegido podría basarse en el propósito de una tecnología y, por lo tanto, en la conveniencia del algoritmo. Por ejemplo, si la intención de un cierto algoritmo es aumentar la protección de los niños de cierta edad en un algoritmo para detectar llamadas que reportan abuso doméstico, entonces los niños de esa edad constituyen un grupo protegido para el propósito del análisis de sesgo algorítmico (Chouldechova et al, 2017).

El tercer paso determina el **conjunto de métricas que se utilizarán para el análisis**. En general, estas métricas cuantifican la medida en que un algoritmo trata a las personas de manera diferente (*disparate treatment*) y la medida en que un algoritmo tiene un impacto distinto en diferentes personas (*disparate impact*). Existen múltiples y, a menudo, superpuestas definiciones de métricas que deberían usarse para evaluar el sesgo algorítmico. Sin embargo, debe mantenerse un cierto grado de acuerdo entre las definiciones en cuestión.

PROBLEMATIZACIÓN, HIPÓTESIS DE TRABAJO Y MÉTRICAS

A pesar de la extensión de los sistemas de inteligencia artificial en el ámbito sanitario, sus implicaciones en términos de transparencia y justicia algorítmica han sido poco estudiados (Sendak et al., 2020). Además, los posibles efectos indeseados de las nuevas tecnologías en el ámbito sanitario han sido destacados (Ash et al., 2004).

En primer lugar, se ha revelado que ciertos factores, predictores de riesgo o bases teóricas para el diagnóstico clínico, pueden pasar desapercibidos al fundamentar el análisis médico en los resultados ofrecidos por sistemas basados en inteligencia artificial (Caruana et al., 2015).

En segundo lugar, se ha indicado que en ciertos casos la introducción de *machine learning* ha dado lugar a la **reducción de las capacidades constatadas del personal sanitario** en la toma de decisiones (Hoff, 2011; Tsai et al., 2003).

En tercer lugar, y como ya se ha señalado, cabe tener en cuenta que los algoritmos automáticos de aprendizaje pueden aprender a **predecir riesgos o asignar beneficios sobre la base de información sesgada contra determinados colectivos sociales.**

Se han identificado sistemas que ofrecen resultados más desventajosos contra las poblaciones pobres o no blancas al evaluar variables como la tasa de readmisión hospitalaria o mortalidad (Joynt Maddox et al., 2019; Lum e Isaac, 2016). Esto ha sucedido en sistemas similares al sistema Laura. Por ejemplo, una auditoría de impacto diferencial en un algoritmo de medición de riesgo encontró que, en una puntuación determinada de riesgo, los pacientes negros estaban considerablemente más enfermos que los pacientes blancos. Se advirtió que remediar esta disparidad en la medición algorítmica de riesgo **aumentaría el porcentaje de pacientes negros que reciben ayuda del 17,7 al 46,5 %.**

El origen de este sesgo estaba en que el algoritmo predecía los costos de atención médica en lugar de la enfermedad, pero la variable atención médica era claramente desigual, lo que afectaba a los pacientes negros (Obermeyer et al., 2019). En otro caso, Di Martino et al. (2019) analizaron dos algoritmos (*Fetal Medicine Foundation* y *BCNatal*) para calcular el riesgo a priori de preeclampsia (basado en el historial médico de factores de riesgo) en cada individuo. Con una tasa fija de falsos positivos de 10 %, los riesgos estimados a priori tanto por la Fetal Medicine Foundation como por los algoritmos BCNatal en una población italiana fueron bastante similares, y ambos resultaron confiables y consistentes. No obstante, los autores también constataron que dicha precisión es menor en el caso de las gestantes que eran **mujeres suramericanas**. Por ello, el análisis de los factores de aprendizaje que podrían predisponer un impacto diferencial por grupo de este tipo debe analizarse en profundidad.

Dados los datos de entrada del sistema de procesamiento automático en el sistema Laura, los tipos de sesgo algorítmico advertidos en sistemas similares y el alcance de la auditoría, se ha estimado **evaluar las categorías sexo biológico y edad en el análisis de impacto algorítmico diferencial**. De este modo, será posible establecer la capacidad del sistema para generar un impacto específico en los pacientes en función de las predicciones de riesgo para estos grupos. Sobre la base de los resultados de este análisis cuantitativo y del estudio de impacto social se propondrán recomendaciones para monitoreo y futuro estudio.

Entre las distintas métricas existentes, las siguientes métricas son particularmente relevantes para el caso de Laura:

M1. La precisión medida de manera adecuada para cada uno de los grupos protegidos.

M2. Las tasas de falsos positivos y/o falsos negativos entre los grupos, que deberían ser similares si quiere afirmarse que no existe discriminación en el funcionamiento del algoritmo (es decir, aplicar el criterio de separación para evitar un maltrato desigual, como describen Zafar et al., 2017). En el caso en que recibir un resultado negativo crea una ventaja (sistemas de asistencia como Laura), la tasa de falsos negativos para el grupo protegido (por ejemplo, en un sistema para predecir riesgo de deterioro clínico), el porcentaje de individuos que eventualmente sufren deterioro clínico, pero que por error fueron categorizados como de bajo riesgo, debe ser similar entre distintos grupos.

5. RESULTADOS DEL ESTUDIO OPERATIVO Y DE ACEPTABILIDAD

El personal hospitalario ha visto con sospecha los sistemas basados en técnicas de aprendizaje automático, debido a su frecuente falta de **explicabilidad y transparencia**. Por este motivo, ciertos sistemas algorítmicos en este ámbito se han valorado positivamente por la manera como combinan una alta precisión con una presentación de los resultados que **facilita su interpretación** (Churpek et al., 2016).

Este apartado refleja un primer análisis de la implantación del sistema Laura y su aceptabilidad, basado en los datos recabados hasta el momento sobre estas variables.

ALINEAMIENTO DEL SISTEMA EN EL CENTRO HOSPITALARIO

Como ya se mencionó, durante la primera mitad de 2018 se realizó un estudio de campo con observación directa y entrevistas semiestructuradas para conocer el proceso de trabajo de enfermeras que emplearon la primera versión del sistema Laura (Gonçalves et al., 2020). El análisis señaló que la participación de enfermeras comienza en la fase de desarrollo del sistema (fase de preimplementación), cuando comparten conocimientos científicos, teóricos y prácticos de salud, lo que ha demostrado ser clave para una buena adopción tecnológica.

En esta línea, el proceso de adopción de Laura se compone de diversas fases de formación de recursos humanos y adaptación tecnológica. La empresa a cargo de este sistema tiene un equipo dedicado a la orientación de los centros hospitalarios en los procesos de integración, adaptación y utilización de la herramienta.

En nuestra entrevista con responsables de esta organización (entrevista online, 18-03-2021), se señaló que esto supone un soporte asistencial por parte de personal que tenga experiencia clínica. Para ello, se realizan diferentes formaciones con los equipos a cargo de Laura. Usualmente, dichos equipos están integrados por directivos médicos, asistenciales, personal de enfermería y médicos. Se presenta la herramienta a los equipos y los directivos colaboran en el proceso de implantación. Inicialmente, se introducen las fases de la adopción tecnológica y se explican los protocolos de adaptación del sistema a las necesidades y características del hospital. Luego, se llevan a cabo dos fases de trabajo: el alineamiento y la implantación generalizada.

Figura 5. Proceso de formación con el sistema Laura

Fuente: Laura.

ALINEAMIENTO

El proceso de alineamiento se extiende durante toda la integración del sistema y comienza por una o dos áreas específicas del centro asistencial. El proceso se compone de una parte técnica, dirigida por el equipo técnico de Laura, y de un alineamiento asistencial, donde participan expertos de la empresa, personal de enfermería y coordinadores del centro hospitalario.

Este proceso consta de:

A. **Alineamiento asistencial:** en esta fase se realiza un diagnóstico asistencial y clínico colaborativo, donde se tratan los procesos clínicos del centro y sus especificidades rutinarias (por ejemplo, los puntos de entrada de información clínica), lo que servirá de base a la implantación del sistema. Del conjunto del proceso se deriva el establecimiento de un protocolo interno de actuación.

B. **Alineamiento técnico:** esta fase supone la revisión de las bases de datos del centro, sus sistemas e infraestructura y el ajuste del modelo algorítmico en función del estudio de su eficiencia aplicada a la institución. Como parte de esta fase piloto del sistema, también se desarrolla el entrenamiento en el uso de la plataforma y sus diferentes pasos de validación técnica.

IMPLANTACIÓN GENERALIZADA DEL SISTEMA

En un segundo momento, como parte de la ampliación del uso de la herramienta a otras partes del centro hospitalario, la empresa realiza nuevas formaciones de los equipos clínicos y nuevos eventos de validación técnica. Pasada esta etapa de ampliación se llevan a cabo entrenamientos de personal que, en muchos casos, incluyen más personal del hospital. Actualmente, el conjunto de las fases de implementación no se refleja todavía en un Manual del sistema para el centro hospitalario.

USABILIDAD

Laura ofrece diversos productos como parte del servicio dedicado a la evaluación del riesgo de deterioro clínico: el reporte continuo en pantalla, la plataforma web que ofrece datos en tiempo real y la nueva aplicación móvil, denominada Laura Assistant.

PLATAFORMA WEB

Como puede advertirse en la siguiente imagen, la plataforma web ofrece diversos instrumentos que van más allá del reporte de riesgo de deterioro clínico de los pacientes. Esto incluye la cantidad de pacientes internados y monitoreados (Atención activa), las alertas existentes, es decir, aquellos pacientes con riesgo medido de deterioro clínico (alertas activas), y el tiempo medio en que se introducen datos clínicos sobre los pacientes (tiempo promedio de entrada

de datos (TMED), que incluye todos los datos (evolución, exámenes, medicamentos, etc.) de los pacientes de la institución y no solo el ingreso de constantes vitales.

Figura 6. Interfaz del sistema Laura

Fuente: Laura.

Además, el sistema permite al personal autorizado buscar información específica sobre cada paciente y también datos estadísticos relevantes para la atención médica.

PANEL DE VISUALIZACIÓN

En cuanto a la visualización en los paneles del sistema Laura, el Panel de control general (ver Figura 7) indica el estado clínico, grado de criticidad y alertas de los pacientes. Puede verse una pantalla de TV o de computador. La gravedad de la situación de los pacientes se expresa mediante los colores rojo (alto riesgo) y amarillo (riesgo intermedio). El color azul, por su parte, señala un sector sin pacientes en alerta o en observación.

Figura 7. Panel del sistema Laura

Fuente: Laura.

- El **paciente en el centro del panel** corresponde al paciente más crítico del sector en un momento dado, por lo que debería visitarse primero.
- Mientras los **pacientes con criticidad amarilla** necesitan ser reevaluados por el equipo cada tres horas, aquellos con criticidad **roja** deberían ser reevaluados por el equipo cada hora. Estos tiempos se configuran por cada centro hospitalario. El sistema registra el momento de reevaluación del paciente solo mediante la introducción de signos vitales en el PEP de la institución.
- La parte inferior de la pantalla presenta **cada sector** de centro y su número de pacientes en alerta, atendiendo a un orden vinculado al número de pacientes críticos. La información de cada sector puede revisarse al detalle al hacer clic sobre el mismo.
- El marcador que se encuentra junto a cada cama en alerta indica la **fecha y hora de la última actualización** de alerta para el paciente. Por lo tanto, señala cuándo se inició o actualizó la alerta en función de las nuevas imputaciones de datos.
- Con el fin de ilustrar el motivo de la alerta se utilizan tres indicadores:
    - **un corazón**, que indica "signos vitales", cuando la razón de la alerta se relaciona con alteraciones en los mismos. Implica que el equipo debe verificar dichos signos, seguir los procedimientos clínicos adaptados a la situación e ingresar nuevos datos en el sistema.
    - **un recipiente** para las "pruebas de laboratorio", que indica que dichas pruebas han cambiado, lo que debería dar lugar a su análisis por el equipo médico.
    - **un reloj** para señalar una alerta de "TMED emergente", el cual indica que además de los cambios en los signos vitales, el paciente no fue reevaluado en el tiempo establecido para su criticidad. Esto debería dar lugar a su reevaluación y al reingreso de datos clínicos.

La esquina izquierda del panel de visualización muestra a los pacientes en alerta o que exhiben variaciones significativas en sus signos vitales. Una vez que se actualizan los datos de estos pacientes mediante la modificación de su historia clínica, pasan al costado derecho del panel, como pacientes en "observación". Para que el sistema clasifique dichos pacientes como reevaluados debe incluir al menos tres signos vitales (temperatura, frecuencia cardíaca y frecuencia respiratoria) en su historia clínica.

Todos estos parámetros, sistemas gráficos y datos los describe en forma clara e ilustrativa el **Manual de instrucciones** del sistema Laura.

Los elementos que la literatura identifica como claves para la usabilidad de sistemas de presentación de información EHR, como **la inteligibilidad de las interfaces, los diseños del soporte de la información no confusos y la iconografía coherente e intuitiva** (Raj et al., 2015) parecen estar debidamente considerados en la pantalla del sistema Laura.

APLICACIÓN LAURA ASSISTANT

La **Aplicación Laura Assistant** permite que los hospitales —utilicen o no historias clínicas electrónicas— recolecten de manera fácil los signos vitales de pacientes hospitalizados y que estos datos los procese en tiempo real el algoritmo del sistema Laura para detectar sepsis y deterioro clínico. Todos los datos antes mencionados, junto con otros como resultados de análisis o notificaciones sobre los pacientes, los ofrece esta aplicación.

Figura 8. Aplicación del sistema Laura

Fuente: Laura.

ACEPTABILIDAD EN LAURA

El estudio cualitativo realizado por Gonçalves et al. (2020) reveló una significativa aceptabilidad del sistema por parte del personal de enfermería, vinculado no solo a la utilidad del sistema, sino a su capacidad de transformar las dinámicas de trabajo mediante la provisión de información en tiempo real. Además, Laura ha realizado encuestas informales con los usuarios finales del sistema que sugieren que el sistema **no aumenta su sobrecarga laboral**, pero también que los médicos no **suelen interactuar de forma regular y activa** con la pantalla informativa (entrevista online, 18-03-2021). El desarrollo actual de la aplicación móvil, con una interfaz que tiene una importante cantidad de información sobre la evolución de los pacientes, se orienta a solventar esta deficiencia y a fomentar su toma de decisiones informada.

Todavía **no se han realizado estudios sistemáticos sobre la usabilidad** del sistema. Mediante los sondeos realizados por el equipo de Laura se advierte que el personal hospitalario pone

particular atención en la información relacionada con los pacientes críticos. Estos pacientes serían los que llaman más la atención del personal médico para realizar una evaluación inmediata (entrevista online, 18-03-2021). Destacar su condición de riesgo en la visualización de los datos es algo personal esperado. En esta línea, el sistema pone el foco en la cama (*leito*) con mayor riesgo y ofrece diferentes mediciones señalando las camas sensibles para las diferentes áreas del hospital (ver Figura 7). Las pantallas están ubicadas en estas diferentes zonas del hospital para que el personal de cada sector pueda identificar pacientes que tengan alertas en el sistema Laura. La reacción y atención a las alertas varía en función de cada institución y equipo médico (algunos han indicado un uso menor), pero normalmente el personal debe **incorporar información en el sistema** sobre la situación del paciente después de haberlo evaluado. Uno de los elementos que contribuye al buen funcionamiento del sistema es la capacidad y velocidad de cada equipo para integrar estos datos, evaluada por los estudios mencionados en la sección 2. Esto puede tener un impacto significativo en la **comunicación intrahospitalaria**, dado que es necesario trabajar con el panel, sobre la base de datos actualizados en tiempo real. Asimismo, esto puede permitir un buen traspaso de información entre los diferentes turnos de atención.

GENERACIÓN DEL CONOCIMIENTO CLÍNICO Y TRANSPARENCIA

El equipo de Laura ha señalado que la existencia de un mecanismo informático de sistematización y clasificación sin aprendizaje automático ha favorecido un **mayor conocimiento del rendimiento hospitalario** y que contribuyó a la **reducción de la mortalidad** asociada al control de pacientes críticos. La combinación de una gestión organizada y abierta de la información, sumada a notificaciones de atención o riesgo, con o sin aprendizaje automático, puede mejorar los procesos de toma de decisiones. Además, el proceso de investigación e implantación del sistema Laura ha dado lugar a la identificación de distintas deficiencias en los centros de atención clínica, como la falta de guías y protocolos bien estructurados para la detección de sepsis o deterioro clínico. Así, la relación entre Laura y el hospital se enmarca también en un proceso de transferencia de conocimiento que da lugar a la mejora de estos procesos (entrevista online, 18-03-2021).

Del mismo modo, mediante la definición colaborativa de los indicadores para deterioro clínico, Laura podría contribuir a solventar uno de los problemas identificados por la literatura, que es la falta de consenso sobre cómo diagnosticar la aparición de dicho deterioro en el sector pediátrico, donde los indicadores más utilizados incluyen necesidad de hospitalización o traslado a la UCI (Bradman et al., 2014; Tucker et al., 2009; Miranda et al., 2020).

EXPLICABILIDAD ALGORÍTMICA

Cabe tener en cuenta que, según la información provista por el equipo de Laura (entrevista online, 18-03-2021) los responsables técnicos y asistenciales de los hospitales son **informados sobre la precisión de Laura** a partir de presentarles su tasa de falsos positivos, su sensibilidad/especificidad, *recall* y su matriz de confusión tras el primer modelamiento. De esta forma, se busca no solo adaptar el sistema de inteligencia artificial a las características y necesidades de cada centro, sino también generar conciencia sobre el alcance del sistema e informar el proceso de toma de decisiones, también con respecto a la versión implementada del modelo algorítmico. Por ejemplo, como parte de este proceso las autoridades sanitarias pueden proponer un balance específico entre la sensibilidad y la especificidad del sistema en este proceso. Además, Laura brinda a los centros reportes semanales y mensuales de su funcionamiento, que han demostrado facilitar información valiosa a los equipos y promover ajustes en los protocolos médicos de actuación (entrevista online, 18-03-2021).

RESUMEN DEL ANÁLISIS DE ACEPTABILIDAD Y RECOMENDACIONES ASOCIADAS

Tabla 2. Síntesis del análisis de aceptabilidad y recomendaciones asociadas

Dimensión	Análisis	Recomendaciones
Usabilidad	Se advierte claridad y coherencia en términos de inteligibilidad de las interfaces, el diseño del soporte de la información y la iconografía de los sistemas.	Se recomienda realizar encuestas interhospitalarias para establecer las limitaciones en relación con estas variables: inteligibilidad, claridad y coherencia e incorporar resultados a la formación del personal (incluidos los materiales de soporte, como el Manual del Usuario) y al diseño tecnológico.
Aceptabilidad	Tanto la literatura que aborda el sistema como las entrevistas sugieren una buena recepción del sistema por parte los y las usuarias. Además, el personal hospitalario indica que el sistema Laura contribuye a mejorar el rendimiento clínico de estos equipos. No obstante, también se utiliza el panel de información en forma desigual y el impacto del sistema en la actualización de los datos del paciente no es siempre significativo, como revela el trabajo de Kalil et al. (2018).	Se recomienda testear la frecuencia en la utilización del sistema en diferentes hospitales y áreas de atención clínica, tanto en términos de tiempos como mediante indicadores de rendimiento en la detección y mitigación de riesgo de deterioro clínico. En este contexto, también se sugiere evaluar el impacto de la utilización de Laura Assistant en términos de la interacción médico-máquina y la introducción de información sobre el paciente (signos vitales). Sobre esta base deberían establecerse mecanismos como la formación o la mejora de los manuales, que permitan resolver posibles problemas de rendimiento general y departamental.

Generación de conocimiento y transparencia	La utilización del sistema ha permitido una mayor digitalización de los servicios hospitalarios y la generación de conocimiento sobre el rendimiento clínico. Dicha información sirve tanto a fines de la mejora del sistema de detección de riesgo como para la atención médica en casos de pacientes con bajo riesgo de deterioro clínico. Por otro lado, la generación de datos no se limita a los registros hospitalarios, sino que también incluye los resultados del procesamiento algorítmico. Dicho conocimiento es compartido en forma abierta y colaborativa con el personal técnico del hospital. No obstante, el conocimiento sobre el alcance y las limitaciones del modelo algorítmico (por ejemplo, de precisión por grupos) no está siendo comunicada en forma exhaustiva al conjunto del personal durante el proceso de alineamiento descrito en la sección anterior.	Se recomienda realizar validaciones regulares sobre la incidencia de Laura en la calidad de los datos clínicos en el registro digital del hospital. Con respecto a la explicabilidad algorítmica se sugiere incorporar en forma sistemática y comprensible para un público general la información (Model card - Mitchell et al., 2019) sobre el funcionamiento del modelo algorítmico, que incluya: 1. Objetivos del sistema; 2. Datos; 3. Aproximación metodológica; 4. Descripción del algoritmo y 5. Parámetros de evaluación de desempeño y errores. Asimismo, dicha información debería ser comunicada, como parte de la política de privacidad del hospital, a todos los pacientes cuyos datos serán objeto de medición por el sistema.

RESULTADOS DEL ANÁLISIS DE ADMINISTRACIÓN DE LOS DATOS PERSONALES

Laura recolecta, y procesa en la nube, datos de las EHR de cada hospital, sobre todo en relación con las variables de análisis antes mencionadas (entrevista online, 17-03-2021). Por lo tanto, los hospitales establecen la conexión al sistema mediante internet. El personal hospitalario y técnico puede acceder al sistema en https://laurabot.laura-br.com/#/access/login utilizando un **sistema de acceso** que consiste en las credenciales de correo electrónico y contraseña.

Figura 9. Autenticación en el sistema Laura

Fuente: Laura.

Como puede verse en la figura anterior, el sistema Laura utiliza un CAPTCHA (*Completely Automated Public Turing test to tell Computers and Humans Apart*: test de Turing público y automático para diferenciar computadores y seres humanos), lo cual es una medida de seguridad para bloquear bots y evitar descifrado de contraseñas.

La **Política de privacidad** de la web Laura (https://www.laura-br.com/politica-de-privacidade.html) incorpora los requerimientos de protección de datos básicos que corresponden a los estándares del Reglamento General de Protección de Datos (GDPR). Esto incluye los objetivos del procesamiento, datos personales implicados, derechos ARCO y datos de contacto del responsable de protección de datos. No obstante, se indica que los datos de los usuarios que dejan comentarios en la web se **conservan indefinidamente con fines de identificación y filtrado de los mismos**, lo cual no se corresponde con el principio de minimización de los datos. Por otro lado, no se ofrece información detallada sobre terceras partes implicadas en el procesamiento de datos personales.

En cuanto a la administración de datos personales por parte del sistema Laura, los registros médicos se procesan en **forma seudonimizada** mediante el uso de las variables de deterioro clínico **asociadas a un identificador no personal por paciente**. Este identificador único sería su número de cama ("*leito*"), que está asociado a su número de registro (entrevista online, 18-03-2021). Asimismo, el registro se utiliza para buscar al paciente en la plataforma Laura.
Los **requerimientos de seguridad de los datos personales** se establecen en Laura a partir del contrato entre la empresa y el centro hospitalario contratante. Sus cláusulas de protección de datos están alineadas con los estándares de la GDPR y la Ley General Brasileña de Protección de Datos (entrevista online, 18-03-2021). Además, las bases de cumplimiento en este ámbito se están alineando actualmente con los requerimientos de la *Health Insurance Portability & Accountability Act*, la principal herramienta legal de protección de los datos correspondientes a los proveedores de servicios sanitarios en Estados Unidos.

Finalmente, **15 de los 40 hospitales** donde el sistema se encuentra operando tienen **Comités de Ética** que pueden evaluar el cumplimiento legal de las investigaciones que lleva adelante Laura con los datos ingresados de los pacientes.

ANÁLISIS Y RECOMENDACIONES

Se ha advertido una tendencia general hacia la aceptabilidad del tratamiento de los datos de los pacientes mediante el uso de EHR por parte del personal hospitalario (Alshahrani et al., 2021). No obstante, cabe considerar que dichos sistemas administran datos sensibles, como el historial médico del paciente, su historial clínico, notas de progreso, medicamentos, signos vitales, inmunizaciones, datos de laboratorio, informes de radiología y datos administrativos. Según la nueva Ley General de Protección de Datos (LGPD) de Brasil (2020), que está alineada

con los requerimientos de la GDPR europea, dicha categoría de datos requiere **recaudos especiales en su tratamiento**. Estas medidas se vinculan fundamentalmente a la necesidad de **confidencialidad, integridad y disponibilidad de la información** en cuestión (Kuturura and Cilliers, 2016). Los principales riesgos a los que se someten los sistemas de información que administran EHR en relación con estos requerimientos incluyen los **ataques** mediante virus o el **acceso no autorizado**, sea intencional o no intencional, por parte del personal hospitalario (Cilliers, 2017:3).

En Laura, la capacidad de asegurar la integridad y confidencialidad de la información se vincula particularmente a su **estrategia de seudonimización de los pacientes** que se encuentran en evaluación de riesgo y monitoreo. Si un conjunto de datos se anonimiza a un nivel alto, es decir, si se eliminan todos los datos personales del historial médico recibido por Laura, su utilidad para terceros disminuye drásticamente. Del mismo modo, cuanto más útil es el conjunto de datos, menos anonimizado suele ser (Ohm, 2009; Lubarsky, 2017).

Dependiendo de la técnica de anonimización, la compensación es diferente. Es decir, cada técnica tiene sus ventajas, deficiencias y problemas asociados. En este caso, se ha optado por la utilización de un sistema de codificación para seudonimizar al paciente y presentar su información en el sistema, utilizando su número de cama como identificador. La seudonimización es una forma de garantizar la identificación y el vínculo continuos con uno o más conjuntos de datos sin identificar directamente a la persona. Normalmente, implica la sustitución de un valor, como un identificador personal, por otro valor. La persona cuyo registro ha sido seudonimizado seguirá siendo identificable debido a la atribución de este nuevo valor. Por ejemplo, João Antunes se convierte en usuario 3849562. Con este sistema, una persona que ha realizado un examen puede buscar el resultado de su prueba en la base de datos con la identificación única que se le dio, sin que otros puedan identificar a una persona específica. Este es un método alternativo al anonimato que a veces resulta suficiente, dependiendo de los datos y sus usos.

Sin embargo, si los cuasi identificadores permanecen dentro del conjunto de datos, el individuo aún es reidentificable. La literatura ha establecido que un seudónimo no es útil para proteger la privacidad si el mismo seudónimo único se usa continuamente en uno o varios conjuntos de datos (Lubarsky, 2017; Article 29 Data Protection Working Party 29, 2014), especialmente cuando la cantidad de atributos vinculados a un registro es elevada y creciente (Barocas y Nissenbaum, 2014), como en el caso del sistema Laura. Las posibilidades de vinculación, singularización e inferencia siguen siendo las mismas entre un conjunto de datos seudonimizados y el conjunto de datos original[15]. Como tal, el Working Party enfatizó fuertemente en su documento de opinión (2014) que un conjunto de datos seudonimizados no se encuentra anonimizado, ni cumple los estándares de anonimización. No obstante, la seudonimización puede utilizarse en combinación con otras técnicas de anonimización con el propósito de anonimizar de forma robusta un conjunto de datos.

Además de estas cuestiones, las medidas que conviene tener en cuenta en el sistema Laura deben incluir:

- Monitoreo de la seguridad en el **mecanismo de acceso al sistema Laura**. La identidad del personal que accede al sistema deberá verificarse mediante un mecanismo de autenticación seguro y que se adapte a los privilegios de cada usuario —niveles de acceso— (Cilliers, 2017). En caso de no usar biometría, se recomienda utilizar protocolos de acceso que combinen contraseñas/códigos de identificación personal con alguna información

15 Este es especialmente el caso si se utiliza un algoritmo predeterminado para seudonimizar un conjunto de datos, como lo explica Lubarsky (2017).

solo conocida por el o la trabajadora (i.e. tarjeta de identificación hospitalaria).

- Para confirmar la **integridad de la información** transmitida mediante el sistema Laura es posible implementar *file hashing*, mediante el uso de un algoritmo de filtrado que utilice el valor de los bits del archivo para confirmar la calidad del mismo y que evite que el documento se vea comprometido (Laudon y Laudon, 2010). Clasificar la sensibilidad de la información, sujeta a distintos niveles de acceso, puede contribuir también a este objetivo.
    - La información debería ser asimismo protegida mediante *firewalls* y monitores de intrusión.
- En cuanto a la **transmisión y almacenamiento**, dado que la información se administra en la nube, se recomienda utilizar encriptación en la comunicación con el sistema Laura. Además, se sugiere establecer algún mecanismo de *non repudiation* para asegurar que se comprueban los datos al ser recibidos por ambas partes y el hospital recibe una prueba de identidad del sistema Laura con respecto a la información enviada como, por ejemplo, las evaluaciones de riesgo realizadas (Maconachy et al., 2001).
    - **Registro continuo** de los accesos (logs) al sistema.
- La **formación de los trabajadores** debe incorporar elementos sobre protección de los datos personales. Esto incluye tipos de datos administrados por Laura, finalidades específicas, requerimientos legales, medidas de protección de la información personal y de transparencia, y comunicación con los pacientes.

Tabla 3. Síntesis del análisis de aceptabilidad y recomendaciones asociadas

Variable	Análisis	Recomendaciones
Política de privacidad	La política de privacidad de la web es completa y está alineada con los requerimientos de la normativa internacional en este campo.	Se sugiere revisar el período de retención de datos personales y seguir el principio de minimización de datos, siempre y cuando su eliminación no afecte la calidad del servicio o el propósito de la tecnología en cuestión.

Seguridad de los datos	El sistema cuenta con un sistema de nombre, contraseña y captcha para el acceso a los datos.	Revisar el sistema de autenticación de la identidad, incorporando claves de acceso con información solo conocida por el personal en cuestión. Integrar un mecanismo de registro y seguimiento de los accesos al sistema, con monitoreo de posibles accesos no autorizados. Asegurar una buena calidad y protección de los datos, mediante file hashing y sistemas de protección frente ataques. Asegurar una debida encriptación en la comunicación y almacenamiento de los datos. Realizar una formación sobre protección de datos a los miembros de Laura y personal hospitalario, abordando los requerimientos básicos de protección de datos sensibles.
Seudonimización y minimización de los datos	El sistema de seudonimización utilizado para la presentación de los datos de deterioro clínico reduce la exposición de datos personales. No obstante, se trata de un sistema de codificación que esté acompañado de otros datos (como número de cama), lo cual facilita la reidentificación de la persona.	Se recomienda revisar la política de seudonimización para garantizar el mayor nivel de confidencialidad y no reidentificación posible en el marco de la funcionalidad requerida.

Fuente: elaboración propia.

6. RESULTADOS DEL ANÁLISIS ALGORÍTMICO

Esta sección presenta una síntesis de los resultados del análisis algorítmico basado en la metodología antes expuesta. El estudio se orienta a evaluar el **impacto diferencial del sistema Laura en los grupos afectados**, focalizando en los atributos sexo biológico y edad, así como sus intersecciones.

Antes de examinar los resultados de la medición de impacto diferencial por grupos se describe la estructura de los datos de "entrada" del algoritmo y también los resultados observados o reales con respecto a la variable deterioro clínico en la población estudiada. Este análisis busca fundamentalmente comprender la base histórica y real de aprendizaje del algoritmo y contrastar sus predicciones de riesgo.

El equipo de Laura ha facilitado un conjunto de datos (dataset) compuesto por 2**874 registros**

hospitalarios, correspondientes a un hospital del sur de Brasil donde el sistema Laura se encuentra en funcionamiento. Los registros se procesaron durante el **año 2020**. La información facilitada comprende un conjunto de datos con esta información: sexo, edad, sector hospitalario, probabilidad de *outcome* (número entre 0 y 1), umbral de *outcome* (número entre 0 y 1), y predicción de *outcome* (número binario, 0: alta hospitalaria, 1: deceso) y *outcome* real (número binario, 0: alta hospitalaria, 1: deceso).

Para analizar el riesgo predicho por el sistema Laura se formateó el conjunto de datos de acuerdo con los requisitos de equidad (*aequitas*); es decir, se renombran las columnas *prediction_outcome* como score y *real_outcome* como *label_value*, creando así mismo tres grupos de análisis basados en los atributos protegidos (sexo, edad y sexo + edad).

ESTRUCTURA SOCIODEMOGRÁFICA

Los datos utilizados en esta auditoría son un conjunto compuesto de 2874 registros hospitalarios antes mencionados. Como esta auditoría se orienta a establecer la eficiencia del sistema en la medición de riesgo por grupos según las variables sexo biológico y edad, la población se ha desagregado en grupos por sexo biológico e intersectado por grupos etarios.

En el conjunto de las 2874 unidades se observa un leve **predominio de la población femenina** sobre la masculina en el número total de pacientes. En segundo lugar, el grupo femenino de pacientes cuenta con una mayor cantidad de registros para los grupos etarios más jóvenes (18-29, 30-39, 40-49 y 50-59); por lo tanto, hay menor cantidad de unidades en los grupos 60-69, 70-69 y 80 o más. Se evidencia que los grupos etarios por **debajo de 17 años tienen muy baja prevalencia**, con 0 registros para el grupo 0-15. Finalmente, la siguiente Tabla muestra también los resultados de "alta hospitalaria" —que alcanzan un 85.4 %— y "deceso" —que acumulan un 14.6 %— del total de la población.

Tabla 4. Representación demográfica de edad, sexo y resultado

Edad	**#**	**%**
0-15	0	-
16-17	7	0.2 %
18-29	119	4.1 %
30-39	241	8.4 %
40-49	435	15.2 %
50-59	567	19.7 %
60-69	775	27.0 %
70-79	528	18.4 %

Sexo	**#**	**%**
Masculino	1350	47 %
Femenino	1524	53 %
Total	**2874**	**100 %**

Resultado	**#**	**%**
Alta hospitalaria	2453	85.4 %
Deceso	421	14.6 %

80+	202	7.0 %

Fuente: elaboración propia.

En el siguiente análisis se decidió eliminar a los pacientes menores de edad, porque su número es muy pequeño (7 individuos en total), lo que hace que el modelo no sea validable para esta población. Además, su valor representa una discontinuidad en la curva de edad, lo que podría indicar que hay circunstancias externas que están influyendo en que este número sea tan pequeño.

Con base en la nueva distribución 18-39, 40-59, 60 o más, la **media de edad de los pacientes** incluidos en el grupo de datos examinado se sitúa cerca de 60 años, aunque es algo más elevada en el caso del sexo masculino. Esta estructura de los datos de entrada es lógica, dado el creciente riesgo de deterioro clínico con el paso de los años y la prevalencia de dicho estado, contrastada por la literatura en la población de sexo masculino.

Figura 10. Número de pacientes por rango de edad

Fuente: Elaboración propia.

ANÁLISIS DE IMPACTO Y TRATAMIENTO DIFERENCIAL POR GRUPOS

A continuación, aparecen los resultados del análisis de impacto diferencial del sistema Laura por grupos, según su sexo. Dicho impacto se analiza mediante tres estrategias. Por un lado, se realiza la comparación entre el riesgo observado y el predicho por ese sistema para el mismo grupo de pacientes. Por otro lado, se miden y examinan las tasas de ratio de predicción negativa (FNR) y ratio de predicción positiva (PPV) para los mismos grupos de edad y sexo. Finalmente, se identifica y analiza la curva de calibración, con el fin de establecer la relación entre valores de predicción y porcentaje de positividad por grupos.

¿QUÉ SE ANALIZA CON EL RATIO DE PREDICCIÓN POSITIVA (PPV)?

El PPV es la tasa de predicción positiva, es decir, el **porcentaje de casos correctamente predichos entre todas las predicciones positivas realizadas**. En el sistema Laura esto representa la proporción de pacientes que experimentaron deterioro clínico en el período recogido, entre los pacientes para los cuales se predijo que eso sucedería.

La **PPVd comparada por grupos (disparidad)** permite, por lo tanto, proyectar si el algoritmo desfavorece a ciertos colectivos ya socialmente desfavorecidos. En el contexto del algoritmo implementado por Laura ello representa que este grupo de pacientes se verá sujeto a tener menos posibilidades de ser atendidos en una situación de deterioro clínico.

En el análisis intragrupal de disparidad de PPVd se toma como referencia que el valor de los PPV sea entre 80 % y 125 % o, lo que es lo mismo, que el PPVd esté entre 0.8 y 1.25.

¿QUÉ SE ANALIZA CON EL RATIO DE FALSOS NEGATIVOS (FNR)?

La tasa de falsos negativos (*False Negative Rate*, FNR) indica la probabilidad del modelo para predecir que un paciente no se encuentra en riesgo de deterioro clínico cuando, en realidad, sí lo está. Esto supondría que un paciente que necesita una ayuda, no la recibiría. **Mayor FNR indica mayor probabilidad de infravaloración del riesgo.** La tasa de falsos positivos (*False Positive Rate*, FPR) indica la probabilidad del modelo para predecir que un paciente se halla en riesgo de deterioro clínico cuando, en realidad, no lo está, lo cual supondría que un paciente que no requiere ayuda, la recibiría, pudiendo dejar sin ella, como consecuencia, a otro paciente que la necesitaría más.

False Negative Rate Disparity **(FNRd) es la proporción de pacientes con un resultado observado conocido ("riesgo de deterioro clínico en el período recogido") para la cual la predicción de ese resultado es de "bajo riesgo" en relación con otros grupos.** En este contexto específico hay interés en tener una tasa baja de falsos negativos. En otras palabras, nos gustaría evitar los casos de pacientes que corren un mayor riesgo de daño (en este caso, de "riesgo de deterioro clínico en el período recogido"), pero que podrían no recibir la atención necesaria, en parte porque el algoritmo predice erradamente el bajo riesgo.

En cuanto al análisis intragrupal de disparidad de FNRd se toma como referencia que el valor de los FNR sea entre 80 % y 125 % o, lo que es lo mismo, que el FNRd esté entre 0.8 y 1.25.

ANÁLISIS DEL IMPACTO DIFERENCIAL POR SEXO

RIESGO OBSERVADO Y RIESGO PREDICHO POR SEXO

La siguiente tabla refleja la distribución de los **resultados reales para los *outcomes* alta y deceso en el conjunto de los datos —habiendo eliminado el grupo de menores de edad—**, tanto en número total como en porcentaje de pacientes para cada grupo según su sexo. Como puede verse, 82.6 % de los pacientes de sexo masculino internados e incorporados en esta auditoría

recibieron el alta en el período analizado. Este número es más elevado para el caso del sexo femenino: 87.7 % de altas en el mismo período. En cuanto al número y **tasa de decesos**, es posible advertir que es **mayor para el sexo masculino, con un 17.4 % del total**. En cambio, **dicha tasa fue de 12.3 % de las 1521 pacientes de sexo femenino**.

Tabla 5. Riesgo observado y riesgo predicho, por sexo

	OBSERVADO				PREDICHO				
Sexo	**Alta #**	**Alta %**	**Deceso #**	**Deceso %**	**Alta #**	**Alta %**	**Deceso #**	**Deceso %**	**Total #**
Masculino	1111	82.6 %	235	17.4 %	926	68.8 %	420	31.2 %	1346
Femenino	1335	87.7 %	186	12.3 %	1123	73.8 %	398	26.2 %	1521

Fuente: elaboración propia.

Al analizar los **resultados predichos por sexo** se observa que el sistema asigna 73.8 % de altas para el grupo de sexo femenino sobre el total para este grupo, y 68.8 % para el masculino, es decir, un número menor. En sentido opuesto, los decesos son mayores en hombres, tanto en número como en porcentaje.

Si bien los datos predichos siguen una misma tendencia en la distribución de riesgo por sexo que en los resultados reales —asigna menos riesgo de deceso al grupo de sexo femenino—, **cabe tener en cuenta que el riesgo predicho en estos resultados supera ampliamente al identificado en los datos observados**. Para el riesgo de deceso en el grupo de sexo masculino, dicha distancia es de **31.2 % predicho frente a 17.4 % observado** y, en el caso de sexo femenino, de **26.2 % predicho frente a 12.3 % observado**. De este modo, la predicción duplica y triplica la tasa de decesos reales, respectivamente.

PREDICCIÓN POSITIVA POR SEXO

Como se observa a continuación, el algoritmo predice adecuadamente el riesgo de deterioro clínico en torno a 50 % de las veces para ambos grupos en los casos identificados como de riesgo. Además, el porcentaje de riesgo predicho correctamente (*true positive*) entre todas las predicciones positivas realizadas por el robot Laura es **menor para el sexo femenino que para el masculino**, si bien dicha diferencia general que afecta al grupo protegido (femenino) es relativamente baja.

Tabla 6. Predicción positiva por sexo

Sexo	**Verdaderos Positivos**	**Falsos Positivos**	**PPV**	**PPV Disparity**
Masculino	223	197	53.1%	1.21
Femenino	175	223	44.0%	1.00

Fuente: elaboración propia.

TASAS DE FALSOS NEGATIVOS POR SEXO

Como puede observarse en la Tabla 7, la variación en las Tasas de Falsos Negativos (FNR) es muy baja y la disparidad entre estas tasas por grupos de sexo masculino y femenino es inferior a 15 %. Esto implica que el sistema tiende a subestimar poco frecuentemente el riesgo de deterioro clínico y que la frecuencia de dicha **subestimación es casi igual para ambos grupos, aunque**

más elevada para el sexo femenino. Cabe tener en cuenta que este último grupo es mayor en número total de unidades en el conjunto de datos y también en el porcentaje de altas, lo que podría explicar esta tendencia.

Tabla 7. Falsos negativos por sexo

Sexo	Verdaderos Positivos	Falsos Negativos	FNR	FNR Disparity
Masculino	223	12	5.1 %	0.86
Femenino	175	11	5.9 %	1.00

Fuente: elaboración propia.

ANÁLISIS DE IMPACTO DIFERENCIAL POR EDAD

RIESGO OBSERVADO Y RIESGO PREDICHO POR EDAD

Al analizar los outcomes recogidos por el sistema Laura por grupos etarios se advierte que las **mayores tasas de altas se producen en los grupos entre 18 y 39 años**, y que dicha tasa es decreciente en cada grupo hasta 60 años o más. Por otra parte, las **mayores tasas de decesos se dan en los grupos entre 40 o más**, siendo la franja 60 o más donde se sitúa la mayoría de los decesos. En cuanto a los números totales, el grupo entre 60 y 69 años acumula el mayor número de pacientes con alta y con deceso.

Tabla 8. Riesgo observado y riesgo predicho por edad

	OBSERVADO				PREDICHO				
Edad	Alta #	Alta %	Deceso #	Deceso %	Alta #	Alta %	Deceso #	Deceso %	Total #
18-39	328	91.1 %	32	8.9 %	304	84.4 %	56	15.6 %	360
40-59	890	88.8 %	112	11.2 %	786	78.4 %	216	21.6 %	1002
60+	1228	81.6 %	277	18.4 %	959	63.7 %	546	36.3 %	1505

Fuente: elaboración propia.

Los datos predichos también reflejan una distribución en la cual el **mayor porcentaje de altas se condensa en los grupos entre 18 y 39 años** y de decesos de 60 o más años. Es posible advertir cómo el porcentaje predicho de decesos crece en las predicciones del sistema Laura a partir de 50 años, pero **prácticamente duplica en todas las franjas al porcentaje observado**.

PREDICCIÓN POSITIVA POR EDAD

Al analizar la proporción de pacientes con riesgo de deterioro clínico, por grupos de edad y en el período recogido, para la cual la predicción de ese resultado entre los resultados positivos ha sido porcentualmente mayor, se advierte una **tasa levemente más elevada para el grupo entre 18 y 39 años**, mientras que para el resto de los grupos las diferencias en PPV son menores a 5 % y la disparidad por grupos (PPVD) muy baja.

Una mayor tasa de positivos entre los pacientes con riesgo predicho, entre 18 y 39 años, no se debería tanto a sus tasas de deterioro real, sino debido a que es el grupo con menor cantidad de pacientes.

Tabla 9. Predicción positiva por edad

Edad	Verdaderos Positivos	Falsos Positivos	PPV	PPV Disparity
18-39	30	26	53.6 %	1.12
40-59	106	110	49.1 %	1.03
60+	262	284	48.0 %	1.00

Fuente: elaboración propia.

TASAS DE FALSOS NEGATIVOS POR EDAD

Entre los pacientes con riesgo de deterioro clínico, **Laura tiende a subestimar este riesgo más a menudo en los pacientes entre 18 y 39 años**. Dicho grupo coincide asimismo con aquellos que cuentan con menor cantidad de pacientes, lo que podría indicar un sesgo derivado de la composición en los datos de entrada. En cambio, en el caso del grupo de 60 o más años, el riesgo de deterioro clínico tiende a sobrestimarse en forma levemente más frecuente que en el resto de grupos etarios.

Tabla 10. Falsos negativos por edad

Edad	Verdaderos Positivos	Falsos Negativos	FNR	FNR Disparity
18-39	30	2	6.2 %	1.15
40-59	106	6	5.4 %	0.98
60+	262	15	5.4 %	1.00

Fuente: elaboración propia.

ANÁLISIS DE IMPACTO DIFERENCIAL INTERSECTADO POR GRUPO ETARIO Y SEXO

RIESGO OBSERVADO Y RIESGO PREDICHO POR EDAD Y SEXO

Como puede observarse en la Tabla 11, **el riesgo de deceso predicho es más elevado que el observado**, una diferencia que se acrecienta con la edad de los pacientes. Las altas observadas superan a las predichas en todos los casos, pero por una diferencia no significativa en términos de tratamiento diferencial. No obstante, mientras la tasa de riesgo de deceso predicho por el sistema Laura es muy similar entre los sexos masculino y femenino para la franja de 60 años o más, esta diferencia tiende a ampliarse entre ambos grupos en las franjas de menor edad.

Tabla 11. Riesgo observado y riesgo predicho por edad

	OBSERVADO	PREDICHO

Edad	Sexo	Alta #	Alta %	Deceso #	Deceso %	Alta #	Alta %	Deceso #	Deceso %	Total #
18-39	M	123	86.0 %	20	14.0 %	115	80.4 %	28	19.6 %	143
18-39	F	205	**94.5 %**	12	5.50 %	189	87.1 %	28	12.9 %	217
40-59	M	336	84.8 %	60	15.2 %	294	74.2 %	102	25.8 %	396
40-59	F	554	91.4 %	52	8.6 %	492	81.2 %	114	18.8 %	606
60+	M	652	80.8 %	155	**19.2 %**	517	64.1 %	290	35.9 %	807
60+	F	576	82.5 %	122	17.5 %	442	63.3 %	256	**36.7 %**	698

Fuente: elaboración propia.

PREDICCIÓN POSITIVA POR EDAD Y SEXO

La Tabla 12 muestra la proporción de casos de riesgo positivo identificados por el sistema Laura en el conjunto de casos positivos (PPV). En línea con los resultados antes expuestos, el PPV es mayor **para el grupo de sexo masculino entre 18 y 39 años que para el femenino en la misma franja de edad**. Esto se evidencia en el número de falsos positivos por grupo de edad, donde el sistema muestra mayor ratio de PP para el grupo protegido (femenino). También en la tasa de predicción positiva, que es muy elevada para el sexo masculino y casi 30 % más baja para las mujeres. En cambio, dicha disparidad se reduce consecutivamente por grupos etarios.

Tabla 12. Predicción positiva por edad y sexo

Idade	Sexo	Verdaderos Positivos	Flasos Positivos	PPV	PPV Disparity
18-39 anos	Masculino	19	9	67,9 %	**1,35**
18-39 anos	Femenino	11	17	39,3 %	**0,78**
40-59 anos	Masculino	58	44	56,9 %	1,13
40-59 anos	Femenino	48	66	42,1 %	0,84
60 o mias	Masculino	146	144	50,3 %	1,00
60 o mais	Femenino	116	140	45,3 %	0,90

Fuente: elaboración propia.

La **diferencia de 0.78 a 1.35 en PPVd** implica que el riesgo de deterioro clínico de una cierta cantidad de pacientes de sexo femenino entre 18 y 39 años podría estar siendo infravalorado en forma frecuente y diferencial. Esto podría explicarse debido a tres factores. Primero, por la **cantidad de pacientes** de sexo femenino (217), que es mayor a la de sexo masculino (143) en esta franja. Sin embargo, cabe tener en cuenta que dicha diferencia es mayor para el grupo 40-59 (396 hombres y 606 mujeres), pero la disparidad en la ratio de predicción positiva se reduce. En segundo lugar, otra explicación es la **elevada tasa de altas que presenta el sexo femenino en esta franja (95 %)**. En tercer lugar, podría vincularse con los **datos clínicos procesados por el algoritmo como predictores de riesgo** (saturación de oxígeno, ratio de respiración, nivel de glucosa en sangre o presión arterial) que podrían reflejar mejores condiciones clínicas para el sexo femenino situado en este grupo etario en forma estadísticamente significativa.

TASAS DE FALSOS NEGATIVOS POR EDAD Y SEXO

Finalmente, en línea con lo anterior, la Tabla 13 muestra cómo el sistema Laura tiende a

subestimar el riesgo de deterioro clínico más a menudo en pacientes de sexo femenino entre 18 y 59 años, que en los de sexo masculino. Notoriamente, esta diferencia también se advierte en el caso de las mujeres en el grupo de 40 a 59 años. Esto cuestiona una explicación del sesgo basada en el número de pacientes procesados. Del mismo modo, debilita una explicación de las diferencias advertidas entre los grupos de sexo masculino y femenino en el grupo de 18 a 39, debida a los mejores datos clínicos observados en el sexo femenino.

Tabla 13. Falsos negativos por edad y sexo

Edad	Sexo	Verdaderos positivos	Falsos Negativos	FNR	FNR Disparity
18-39 años	Masculino	19	1	5.0 %	0.86
18-39 años	Femenino	11	1	8.3 %	1.43
40-59 años	Masculino	58	2	3.3 %	0.57
40-59 años	Femenino	48	4	7.7 %	1.32
60 o más	Masculino	146	9	5.8 %	1.00
60 o más	Femenino	116	6	4.9 %	0.85

Fuente: elaboración propia.

ANÁLISIS DE LA FUNCIÓN DE *SCORING*

El análisis de calibración se basa en una definición inicial con respecto a la función de scoring. Se trata de una distribución bimodal, entre 0 y 0.3 y entre 0.7 y 1.0. El corte de decisión en 0.069 (línea vertical en la figura 11) **es un corte *arbitrario* que obedece a las características del equipo** y su capacidad de respuesta más que a un riesgo extremo o riesgo de deceso.

El paciente que obtiene un riesgo 0.068 tiene un riesgo similar a alguien que obtiene un riesgo 0.070, aunque su interpretación binaria sea muy diferente. Esta característica debe explicársele al equipo médico que atiende Laura.

Figura 11. Curva de densidad de probabilidad

Fuente: Elaboración propia.

ANÁLISIS DE LA CALIBRACIÓN

Una vez definida la curva de densidad de probabilidad, se realiza un **análisis de las curvas**

de calibración, lo cual permite analizar la probabilidad de deceso en diferentes grupos. En particular, la curva de calibración indica:

- x = puntuación
- y = probabilidad de muerte para las personas con esa puntuación (número de personas que fallecen dividido entre el número de personas totales, dado un rango de puntuación).

Se han calculado deciles sobre la puntuación para calcular los rangos sobre la misma y se han tomado estos umbrales como punto de medición. **La "curva" resultante debería ser una línea recta e igual para diferentes grupo**s (M, F, M + *Young*, M + Old, F + *Young*, F + *Old*).

Como puede advertirse en la figura 12, la curva de calibración es más ajustada a la calibración perfecta en los casos de menor y mayor puntaje. La c**alibración se pierde cerca del punto medio**, donde también se **aprecian mayores diferencias entre hombres y mujeres.** En todos los casos, el puntaje parece sobreestimar el riesgo, es decir, que un puntaje de 0.05, por ejemplo, corresponde a un riesgo menor a 5 %, independientemente del género. Se recomienda buscar una mayor calibración del modelo, en particular en torno al valor que se usa como punto de corte.

Figura 12. Curva de calibración por sexo

Al analizar dichas curvas por grupos etarios se advierten importantes diferencias grupales relacionadas con las medias de riesgo predicho identificadas anteriormente. Por un lado, el grupo de mujeres entre 18 y 39 años evidencia una creciente distancia con respecto a la probabilidad de decesos (descalibración) conforme aumentan los valores de la predicción. El grupo mejor calibrado es el de sexo masculino entre 18-39, a pesar de que comparte una tasa de decesos observada similar al femenino (14 % y 15 %, respectivamente) y un número menor de pacientes. Por otro lado, mientras los grupos de sexo masculino y femenino muestran la misma curva para los grupos masculino 40-59 y 60 o más femenino, el grupo femenino entre 40-59 años presenta también una caída en la positividad desde el valor de predicción 0.2, que se sostiene de forma continua a lo largo de la curva.
Figura 11. Curva de calibración por edad y sexo

CONCLUSIONES Y RECOMENDACIONES

La auditoría del sistema Laura ha abordado diversos aspectos de su diseño y funcionamiento con el fin de establecer una aproximación a sus cualidades en términos de **aceptabilidad, usabilidad, protección de datos y justicia algorítmica**. Con este fin, primero se estableció el estado del arte teórico en torno a los sistemas automáticos de detección de riesgo de deterioro clínico y a la composición del marco social general de implantación. A continuación, se aplicó una serie de estrategias metodológicas y técnicas de recolección de datos orientadas a recoger los datos cualitativos y cuantitativos para la realización del análisis. Estas incluyeron entrevistas semiestructuradas con desarrolladores y personal a cargo de la integración del sistema en el ámbito hospitalario, al igual que la medición de sesgo algorítmico de grupo sobre la base de los falsos negativos y la predicción positiva por grupos.

Cabe señalar que el **análisis de impacto social** se diseñó para brindar una descripción general de los aspectos societales que podrían limitar el funcionamiento del sistema. En esta línea, la auditoría algorítmica se concibió con el fin de identificar **evidencia indirecta de sesgo** sobre la base del estudio de dos variables sociodemográficas fundamentales para el sistema: sexo biológico y edad. Este diseño metodológico, limitado al análisis de unos factores específicos de sesgo, se explica por el alcance de la auditoría acordada con el BID y reflejada en el Plan de Análisis compartido con Laura y adaptado a la situación pandémica actual, que ha limitado el trabajo de campo.

En términos de impacto social, con base en la información recabada, se ha advertido una **significativa inteligibilidad** por parte de la comunicación in situ del sistema, tanto a nivel de organización lógica de la información como a la composición iconográfica. En cuanto a la aceptabilidad de Laura por usuarios finales se han identificado distintas fuentes que indican una **importante aceptación tecnológic**a, tanto por la **facilidad de us**o como por la construcción y transmisión de conocimiento clínico. No obstante, esta buena recepción general la ha **matizado la relativa utilización del sistema** (Kalil et al., 2018), una limitación que busca abordarse mediante el desarrollo de Laura Assistant.

Además, se han identificado **limitaciones** en la transmisión del conocimiento e información a las personas usuarias sobre el alcance y las limitaciones del modelo algorítmico, **particularmente con respecto a la precisión por grupos**, que podría comunicarse de forma más específica. Asimismo, esta cuestión podría ser abordarse con quienes participen en la formación brindada durante el proceso de alineamiento de Laura.

En términos de **protección de datos**, se ha identificado una política de privacidad completa y bien estructurada, un estándar de acceso y autenticación al sistema con mecanismos de seguridad básicos y una política de seudonimización en la transmisión abierta que siguen los principios de privacidad en el diseño. No obstante, se ha sugerido la revisión de estos estándares con el fin de confirmar su proporcionalidad con respecto a la sensibilidad y el volumen de datos personales tratados.

RECOMENDACIONES DEL ANÁLISIS CUALITATIVO

I. Realizar **encuestas interhospitalarias** para establecer las limitaciones en relación con las variables **inteligibilidad, claridad y coherencia**. Incorporar los resultados de estas encuestas en la formación del personal (incluidos los materiales de soporte, como el Manual del Usuario) y el diseño tecnológico.

II. Testear la **frecuencia en la utilización del sistema** en diferentes hospitales y áreas de atención clínica, tanto en términos de tiempos como mediante indicadores de rendimiento en la detección y mitigación de riesgo de deterioro clínico.

    A. En este contexto, también se sugiere evaluar el impacto de la utilización de Laura Assistant en términos de la interacción médico-máquina y la introducción de información sobre el paciente (signos vitales). Sobre esta base deberían establecerse mecanismos como la formación del personal o la mejora de los manuales, de modo que permitan resolver posibles problemas de rendimiento general y departamental.

III. Se recomienda **realizar validaciones regulares sobre la incidencia del sistema Laura en la calidad de los datos clínicos** en el registro digital del hospital.

IV. Con respecto a la explicabilidad algorítmica se sugiere incorporar de forma sistemática y comprensible para un público general la información (*Model card* - Mitchell et al., 2019) sobre el funcionamiento del modelo algorítmico, que incluya: 1. Objetivos del sistema; 2. Datos; 3. Aproximación metodológica; 4. Descripción del algoritmo y 5. Parámetros de evaluación de desempeño y errores.

    A. Dicha información debería comunicarse, como parte de la política de privacidad del hospital, a todos los pacientes cuyos datos serán objeto de medición por el sistema.

V. **Revisar el período de retención de datos personales** y seguir el principio de minimización de datos, siempre y cuando su eliminación no afecte la calidad del servicio o el propósito de la tecnología en cuestión.

    A. Realizar una formación sobre protección de datos a los miembros del sistema Laura y al personal hospitalario, que aborde los requerimientos básicos de protección de datos sensibles.

VI. Examinar la **seguridad del sistema**, que incluya el mecanismo de autenticación de la identidad e incorpore claves de acceso con información solo conocida por el personal en cuestión. Integrar un mecanismo de registro y seguimiento de los accesos al sistema, con monitoreo de posible acceso no autorizado. Asegurar una buena calidad y protección de los datos, mediante file hashing y sistemas de protección frente a ataques. Asegurar una debida encriptación en la comunicación y almacenamiento de los datos.

VII. Revisar la **política de seudonimización** que garantice el mayor nivel de confidencialidad y no reidentificación posible en el marco de la funcionalidad requerida.

En lo referente a la auditoría algorítmica centrada en el análisis de datos, este estudio comenzó analizando 2874 registros hospitalarios que sirvieron de base a la medición de impacto y tratamiento diferencial por grupos. De este modo, se identificaron algunas características del conjunto de datos (dataset), como el predominio de población de sexo femenino, que corresponde a grupos más jóvenes que la de sexo masculino, o la escasa presencia de menores de 17 años, que podrían ayudar a explicar el comportamiento del sistema. Sobre esta base, de hecho, se tomó la decisión de eliminar este grupo menor de registros en el análisis, pues podrían desviar sus resultados.

Luego, se realizaron tres análisis: una comparación entre el riesgo observado y el predicho por Laura para el mismo grupo de pacientes, las tasas de FN y PPV para los mismos grupos de edad y sexo, y la curva de calibración, con el fin de establecer la relación entre valores de predicción y porcentaje de positividad por grupos.

En síntesis, se advierten resultados consistentes con respecto a la menor capacidad predictiva de Laura para las personas del sexo femenino entre 18 y 39 años.

Tabla 14. Resumen de los resultados del análisis de datos

		Resultado
Sexo	Riesgo	El sistema asigna un mayor riesgo de deceso predicho sobre el observado, con una diferencia de 14 puntos porcentuales para ambos sexos. Este margen de riesgo dado por el sistema puede deberse a una calibración que persigue elevar el riesgo predicho, de manera que minimice el riesgo de falsos negativos.
	PPV	El PPV es menor para el sexo femenino que para el masculino; es decir, la probabilidad de que la predicción de deceso sea correcta es menor para mujeres que para varones.
	FNR	Se observan tasas bajas de falsos negativos (~5-6 %). El sistema Laura tiende a subestimar poco frecuentemente el riesgo de deterioro clínico. El FNRd del grupo de sexo masculino es de 0.83, pues el grupo femenino es el que arroja una mayor tasa de falsos negativos.

Edad	Riesgo	Tal y como ocurre en la dimensión sexo, el riesgo predicho es mayor que el riesgo observado. También se observa que el riesgo predicho aumenta en cada grupo etario analizado. El grupo 18-39 años tiene una diferencia de seis puntos porcentuales (pp) entre el riesgo predicho y observado; el grupo 40-59 años tiene 10 pp, y el grupo 60 o más tiene 19 pp. El grupo con menor riesgo observado y predicho es el grupo de 18 a 39 años.
	PPV	El PPV es mayor para el grupo 18-39 años, lo que podría perjudicar a los grupos vulnerables por encima de 70 años. La disparidad por grupos (PPVD) está dentro de los umbrales objetivos, si bien la tasa más elevada de PPVd, que corresponde al grupo 18-39 años, podría explicarse por ser el grupo con menor cantidad de pacientes (360 frente a 1550).
	FNR	Se observan tasas bajas de falsos negativos (~5-6 %); el sistema Laura tiende a subestimar poco frecuentemente el riesgo de deterioro clínico. La disparidad por grupos muestra que Laura tiende a subestimar este riesgo más a menudo para el grupo de 18-39 años.
Sexo y edad	Riesgo	El riesgo predicho aumenta en puntos porcentuales de forma similar a la observada en la dimensión edad. Se observa un mayor incremento en puntos porcentuales para el sexo femenino, es decir, la predicción para el sexo masculino es más cercana a los datos observados que para el sexo femenino.
	PPV	Se observan mayores valores de PPVd en grupos de sexo masculino frente al sexo femenino; se encuentra la mayor discrepancia entre los grupos 18-39 años de sexo masculino (1.35) y 18-39 años de sexo femenino (0.78), estando estos valores fuera de los umbrales objetivo (0.8 - 1.25). Esta diferencia podría explicarse por un sesgo en los datos; el menor riesgo de deterioro clínico observado para el sexo femenino en esta franja etaria podría explicarse por diferencias en las mediciones entre sexos para los datos clínicos de entrada.
	FNR	Se observan rangos de tasas bajas de falsos negativos más amplias (~5-8 %), correspondiendo los dos valores más altos a grupos etarios de sexo femenino. La disparidad por grupos arroja valores por encima del umbral objetivo (1.25) para los grupos de sexo femenino ente 18-39 años (1.43) y 40-59 años (1.32). El sistema Laura tiende a subestimar el riesgo de deterioro clínico más a menudo en estos grupos.
-	Calibración	La curva de calibración es más ajustada a la calibración perfecta en los casos de menor y mayor puntaje, pero se pierde cerca del punto medio, donde también se aprecian mayores diferencias entre hombres y mujeres.

Fuente: elaboración propia.

RECOMENDACIONES DEL ANÁLISIS ALGORÍTMICO

I. Monitorear los grupos con pocos pacientes y eliminar aquellos con muy pocos pacientes, como los menores de 17 años en el conjunto de datos (*dataset*) analizado. En esta línea se recomienda estudiar los **casos de variables con muy baja prevalencia** en la muestra, que se considera que no pueden ser modelables en forma robusta.

   A. Una posibilidad al respecto es incorporar alertas cuando el sistema los detecta.

II. Dado que el sistema tiende a **desproteger a las personas de sexo femenino de entre 18 y 39 años**, se recomienda:

   A. Alertar sobre esta característica a los administradores del sistema. Es decir, advertir al personal hospitalario de que el sistema infraestima el riesgo para este grupo.

III. Se recomienda **buscar una mayor calibración del modelo**, particularmente en torno al valor que se usa como punto de corte. Esto debe asimismo contrastarse con respecto a su efecto en las tasas de PPV y FN en los grupos intersectados (edad y sexo) analizados en este documento.

IV. Garantizar la **preparación necesaria de las y los trabajadores** que interactúen con el modelo durante el proceso de alineación, que incorpore información sobre los márgenes de precisión del mismo para los grupos auditados.

V. **Explicitar** de cara a los y las trabajadoras hospitalarias, así como a los pacientes en general, el objetivo del modelo, aclarando que no se trata de un sistema de decisión autónoma, sino solo de refuerzo objetivo en la toma de decisión.

REFERENCIAS

Alshahrani, A., Jamal, A., and Tharkar, S. (2021). How private are the electronic health records? Family physicians' perspectives towards electronic health records privacy. *Journal of Health Informatics in Developing Countries*, 15(1). Disponible en https://www.jhidc.org/index.php/jhidc/article/view/298

Article 29 Data Protection Working Party. (2014). Opinion 05/2014, on *Anonymisation Techniques*. Disponible en https://www.pdpjournals.com/docs/88197.pdf

Ash, J. S., Berg, M., and Coiera, E. (2004). Some unintended consequences of *information technology in healthcare.* Journal of the American Medical Informatics Association, 11(2): 104-112.

Baeza-Yates, R. (2018). Bias on the web. *Communications of ACM* 61(6): 54-61.

Bandeira da Silva, D., Schmidt, D., da Costa, C. A., da Rosa Righi, R., and Eskofier, B. (2021). DeepSigns: A predictive model based on Deep Learning for the early detection of patient health deterioration. *Expert Systems with Application*s, 165(5). Disponible en https://www.sciencedirect.com/science/article/abs/pii/S0957417420307004

Barocas, S. and Nissenbaum, H. (2014). Big Data's End Run around Anonymity and Consent. In Lane, J., Stodden, V., Bender, S., and Nissenbaum, H. (Eds.), *Privacy, Big Data, and the Public Good: Frameworks for Engagement*, 44-75. Cambridge: Cambridge University Press. doi:10.1017/CBO9781107590205.004

Barocas, S. and Hardt, M. (2017). *Fairness in Machine Learnin*g. Tutorial at NIPS. https://mrtz.org/nips17/

Barocas, S. and Selbst, A. (2016). Big Data's Disparate Impact, *California Law Revie*w, 104: 671-732. Disponible en http://www.californialawreview.org/wp-content/uploads/2016/06/2Barocas-Selbst.pdf

Batista, N. O. W., Coelho, M. C. R., Trugilho, S. M., Pinasco, G. C., Santos, E. F. S., and Ramos-Silva, V. (2015). Clinical-epidemiological profile of hospitalised patients in pediatric intensive care unit. *Journal of Human Growth and Developmen*t, 25(2): 187-193. Disponible en https://dx.doi.org/10.7322/jhgd.103014

Bihorac, A., Ozrazgat-Baslanti, T., Ebadi, A., Motaei, A., Madkour, M., Pardalos, P. M., Lipori, G., Hogan, W. R., Efron, P. A., Moore, F., Moldawer, L. L., Wang, D. Z., Hobson, C. E., Rashidi, P., Li, X., and Momcilovic, P. (2019). MySurgeryRisk: Development and Validation of a Machine-learning Risk Algorithm for Major Complications and Death After Surgery. *Annals of Surgery*, 269(4): 652-662._

Binns, R. (2018). Algorithmic Accountability and Public Reason. *Philosophy & Technology*, 31(4): 543-556.

Bradman, K., Borland, M., and Pascoe, E. (2014). Predicting patient disposition in a pediatric emergency department. *Journal of Paediatrics and Child Health*, 50(10): 39-44. Disponible en

https://dx.doi.org/10.1111/jpc.12011

Cardoso, L. T., Grion, C. M., Matsuo, T., Anami, E. H., Kauss, I. A., Seko, L., and Bonametti, A. M. (2011). Impact of delayed admission to intensive care units on mortality of critically ill patients: A cohort study. *Critical Care*, 15(1): R28.

Caruana, R., Lou, Y., Gehrke, J., et al. (2015). "Intelligible Models for healthcare: predicting pneumonia risk and hospital 30-day readmission". In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Cham, Switzerland: Springer International Publishing AG, 1721-1730.

Castillo, C. (2018). Algorithmic Discrimination. Assessing the impact of machine intelligence on human behaviour: An interdisciplinary endeavour. *Proceedings of HUMAINT Workshop*. Disponible en https://arxiv.org/pdf/1806.03192.pdf

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments, arXiv:1610.07524. Disponible en: https://arxiv.org/abs/1610.07524

Churpek, M. M., Yuen, T. C., Winslow, C., Meltzer, D. O., Kattan, M.W., and Edelson, D. P. (2016). Multicenter Comparison of Machine Learning Methods and Conventional Regression for Predicting Clinical Deterioration on the Wards. *Critical Care Medicine*, 44(2): 368-374.

Cilliers, L. (2017). Exploring information assurance to support electronic health record systems. 2017 *IST-Africa Week Conference (IST-Africa*), Windhoek, Namibia: 1-8. doi: 10.23919/ISTAFRICA.2017.8102363.

Cowgill, B. (2019). Bias and productivity in humans and machines. *Upjohn Institute Working Paper*, No. 19-309, W.E. Upjohn Institute for Employment Research: Kalamazoo. doi: 10.17848/wp19-309. Disponible en: https://research.upjohn.org/up_workingpapers/309/

Danks, D. and John London, A. (2017). Algorithmic bias in autonomous systems. *Proceedings of the 26th International Joint Conference on Artificial Intelligence*. AAAI Press: 4691-4697.

Di Martino, D., Masturzo, B., Paracchini, S., Bracco, B., Cavoretto, P., Prefumo, F., Germano, C., Morano, D., Girlando, F., Giorgione, V., Parpinel, G., Cariello, L., Fusè, F., Candiani, M., Todros, T., Rizzo, N., and Farina, A. (2019). Comparison of two "a priori" risk assessment algorithms for preeclampsia in Italy: A prospective multicenter study. *Archives of Gynecology and Obstetrics,* 299(6): 1587-1596.

Duncan, B. B., Cousin, E., Naghavi, M. et al. (2020). The burden of diabetes and hyperglycemia in Brazil: A global burden of disease study 2017. *Population Health Metrics* 18(9). Disponible en https://doi.org/10.1186/s12963-020-00209-0

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. (2012). "Fairness through awareness". *Proceedings of the 3rd innovations in theoretical computer science conference*: 214-226.

Ferryman, K. and Pitcan, M. (2018). Fairness in precision medicine. *Data & Society*. Disponible en https://datasociety.net/library/fairness-in-precision-medicine/

Gillen, S., Jung, C., Kearns, M., and Roth, A. (2018). *Online learning with an unknown fairness metric*, arXiv:1802.06936. Disponible en https://arxiv.org/abs/1802.06936

Goldstein, B. A, Navar, A. M., Pencina, M. J., and Ioannidis, J. P. A. (2017). Opportunities and challenges in developing risk prediction models with electronic health records data: A systematic review. *Journal of the American Medical Informatics Association*, 24(1): 198-208.

Gonçalves, L. S., Amaro, M. L. M., Romero, A. L. M., Schamne, F. K., Fressatto, J. L., Bezerra, C. W. (2020). Implementation of an Artificial Intelligence Algorithm for Sepsis Detection. *Revista Brasileira de Enfermagen*. 73(3): 1-5.

Green, M., Lander, H., Snyder, A., Hudson, P., Churpek, M., and Edelson, D. (2018). Comparison of the Between the Flags calling criteria to the MEWS, NEWS and the electronic Cardiac Arrest Risk Triage (eCART) score for the identification of deteriorating ward patients. *Resuscitation*, 123: 86-91.

Gulshan, V., Peng, L., Coram, M., et al. (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*, 316(22): 2402-2410.

Haupt, C. E. (2019). Artificial Professional Advice. *Yale Journal of Law Technology*. 21(3): 55-77.

Heidari, H., Ferrari, C., Gummadi, K., and Krause, A. (2018). "Fairness behind a veil of ignorance: A welfare analysis for automated decision making". In: Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N. and Garnett, R. (Eds.). *Advances in Neural Information Processing Systems 31*. Montreal, QC: Curran Associates, Inc.: 1265-1276.

Hoff, T. (2011). Deskilling and adaptation a*mong primary care physicians using two work innovations.* Health Care Management Review, 36(4): 338-348.

Holstein, K., Wortman Vaughan, J., Daumé III, H., Dudik, M., and Wallach, H. (2019). Improving fairness in machine learning systems: What do industry practitioners need? *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Paper No. 600: 1-16.

IBGE (2018). Sistema de Contas Regionais: Brasil 2018. Contas Nacionais, 77. Disponible en https://biblioteca.ibge.gov.br/visualizacao/livros/liv101765_informativo.pdf

Joynt Maddox, K. E., Reidhead, M., Qi, A. C., and Nerenz, D. R. (2019). Association of Stratification by Dual Enrollment Status With Financial Penalties in the Hospital Readmissions Reduction Program. *JAMA Internal Medicine*, 179(6): 769-776.

Kalil, A. J. (2017). Avaliação do impacto na identificação de pacientes com risco de sepse após implantação de um robô cognitivo gerenciador de risco (ROBÔ LAURA®). Dissertação de Mestrado. Curitiba: Universidade Tecnológica Federal do Paraná.

Kalil, A.J., Dias, V. M. C. H., Rocha, C. C., Morales, H. M. P., Fressatto, J. L., and Faria, R. A. (2018). Sepsis risk assessment: A retrospective analysis after a cognitive risk management robot (Robot Laura) implementation in a clinical-surgical unit. R*esearch on Biomedical Engineering*, 34(4): 310-316.

Katurura, M. and Cilliers, L. (2016). "The extent to which the POPI Act makes provision for patient privacy in mobile personal health record systems". In: *The conference proceedings of IST-Africa* 2016, 11-13 May. Durban: IST-Africa.

Kim, M. P., Reingold, O., and Rothblum, G. N. (2018). *Fairness through computationally-bounded awareness*. arXiv:1803.03239. Disponible en https://arxiv.org/abs/1803.03239

Kobylarz Ribeiro, J. et al. (2020). "A Machine Learning Early Warning System: Multicenter Validation in Brazilian Hospitals". *2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS)*. Rochester, MN: 321-326.

Laudon, K. C. and Laudon, J. P. (2010). *Management Information Systems*. New Jersey: Pearson Education.

Lippert-Rasmussen, K. (2013). *Born Free and Equal? A Philosophical Inquiry into the Nature of Discrimination*. Oxford: Oxford University Press.

Loreto, M., Lisboa, T., and Moreira, V. P. (2020). Early prediction of ICU readmissions using classification algorithms. *Computers in Biology and Medicin*e, 118(C).

Lubarsky, B. (2017). Re-identification of "Anonymized Data". *Georgetown Law Technology Review*, 2(1): 202-213.

Lum, K. and Isaac, W. (2016). To predict and serve? *Significanc*e, 13, 14-19.

Maconachy, V. W., Schou, C. D., Ragsdale, D., and Welch, D. (2001). "A model for Information Assurance: An Integrated Approach". In: *The proceedings of the 2001 IEEE Workshop on Information Assurance and Security*. United States Military Academy, West Point, NY, 5-6 June.

Madden, M. (2018). Need medical help? Sorry, not until you sign away your privacy". *MIT Technology Review*, October 23. Disponible en https://www.technologyreview.com/s/612282/need-medical-help-sorry-not-until-you-sign-away-your-privacy/

Miranda, J. O. F. et al. (2020). Factors associated with the clinical deterioration recognized by an early warning pediatric score. *Texto & Contexto Enfermagen*, 29: 1-12.

Mitchell, M. S., Wu, A., Zaldivar, P., Barnes, L., Vasserman, B., Hutchinson, E., Spitzer, I., Raji, D., and Gebru, T. (2019). "Model Cards for Model Reporting". In: Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19). *Association for Computing Machinery,* New York, NY, 220–229. doi:https://doi.org/10.1145/3287560.3287596

Morgan, R., Lloyd-Williams, F., Wright, M., and Morgan-Warren, R. (1997). An early warning scoring system for detecting developing critical illness. *Clinical Intensive Care*, 8: 100.

Muralitharan, S., Nelson, W., Di, S., McGillion, M., Devereaux, P., Barr, N., and Petch, J. (2021). Machine Learning–Based Early Warning Systems for Clinical Deterioration: Systematic Scoping. *Review Journal of Medical Internet Research*, 23(2).

Narayanan, A. (2018). Tutorial: 21 definitions of fairness and their politics *[Abstract and video]. Conference on Fairness, Accountability, and Transparency*. NYC, Feb 23.

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

Ohm, P. (2009). Broken promises of privacy: Responding to the surprising failure of

anonymization. *UCLA Law Review*, 57, 1701-1777. Disponible en https://pages.uoregon.edu/koopman/courses_readings/phil407-net/ohm_broken_promises_privacy.pdf

Pimentel, M. A., Redfern, O. C., Malycha, J., Meredith, P., Prytherch, D. R., Briggs, J., Young, J. D., Clifton, D. A., Tarassenko, L., and Watkinson, P. J. (2021). Detecting deteriorating patients in hospital: Development and validation of a novel scoring system. *Americal Journal of Respiratory and Critical Care Medicine*. Disponible en https://www.atsjournals.org/doi/abs/10.1164/rccm.202007-2700OC

Price, W.N. (2017). Regulating Black-Box Medicine. *Michigan Law Review*, 116(3): 421-474.

Ratwani, R. M., Fairbanks, J. R., Hettinger, A. Z., and Benda, N. C. (2015). Electronic health record usability: Analysis of the user-centered design processes of eleven electronic health record vendors. *Journal of the American Medical Informatics Association*, 22(6): 1179-1182. https://doi.org/10.1093/jamia/ocv050

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1: 206-215.

Sendak, M., Elish, M. C., Gao, M., Futoma, J., Ratliff, W., Nichols, M., Bedoya, A., Balu, S., and O'Brien, C. (2020). "The human body is a black box": Supporting clinical decision-making with deep learning. In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20)*. Association for Computing Machinery, New York, NY, 99-109. doi:https://doi.org/10.1145/3351095.3372827

Speicher, T., Heidari, H., Grgic-Hlaca, N., Gummadi, K. P., Singla, A., Weller, A. and Bilal Zafar, M. (2018). "A Unified Approach to Quantifying Algorithmic Unfairness". Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, July 2018: 2239-2248. doi:10.1145/3219819.3220046

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

Tsai, T. L., Fridsma, D. B., and Gatti, G. (2003). Computer decision support as a source of interpretation error. *Journal of the American Medical Informatics Association*, 10(5): 478-483.

Tucker, K. M., Brewer, T. L., Baker, R. B., Demeritt, B., and Vossmeyer, M. T. (2009). Prospective evaluation of a pediatric inpatient early warning scoring system. *Journal for Specialists in Pediatric Nursing*, 14(2): 79-85. Disponible en https://dx.doi.org/10.1111/j.1744-6155.2008.00178.x

Tume, L. (2007). The deterioration of children in ward areas in a specialist children's hospital. Nursing in Critical Care, 12(1): 12-19. Disponible en https://dx.doi.org/10.1111/j.1478-5153.2006.00195.x

Turney, P. D. (1996). How to shift bias: Lessons from the Baldwin effect. *Evolutionary Computation*, 4(3): 271-295.

Ueno, R., Xu, L., Uegami, W., Matsui, H., Okui, J., Hayashi, H., et al. (2020). Value of laboratory results in addition to vital signs in a machine learning algorithm to predict in-hospital cardiac arrest: A single-center retrospective cohort study. *PLoS ONE*, 15(7): 1-16.

Williams, B. (ed.). (2017). *National Early Warning Score (NEWS) 2 - Standardising the assessment of acute illness severity in the NHS.*

Zfania, T. K., Yang, J.,. Rossetti, S C., Cato, K. D., Kang, M. J., Knaplund, C., Schnock, K. O., Garcia, J. P., Jia, H., Schwartz, J. M., and Zhou, L. (2020). Mining clinical phrases from nursing notes to discover risk factors of patient deterioration. *International Journal of Medical Informatics*, 135. Disponible en https://www.sciencedirect.com/science/article/abs/pii/S1386505619309682