Robot Laura
Auditoría Algorítmica

Estudio sobre el sistema Laura de predicción
de riesgo de deterioro clínico

eticas BID

Robot Laura
Auditoría Algorítmica

Estudio sobre el sistema Laura de predicción
de riesgo de deterioro clínico

Diciembre de 2021

https://www.iadb.org/

Índice

Laura es un **Robot Cognitivo/ Gestor de Riesgos** que actúa en la **identificación temprana de los riesgos de deterioro clínico.**

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 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.

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.

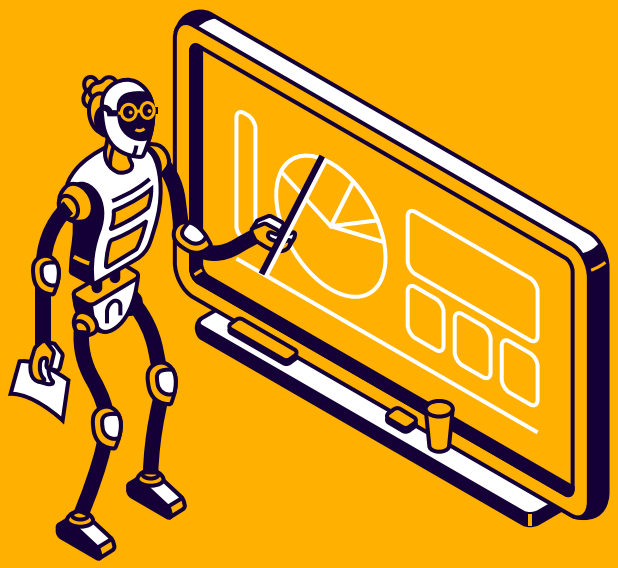
2 Véase presentación en https://www.laura-br.com/en/

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 asistencial del sistema (28-03-2021).

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

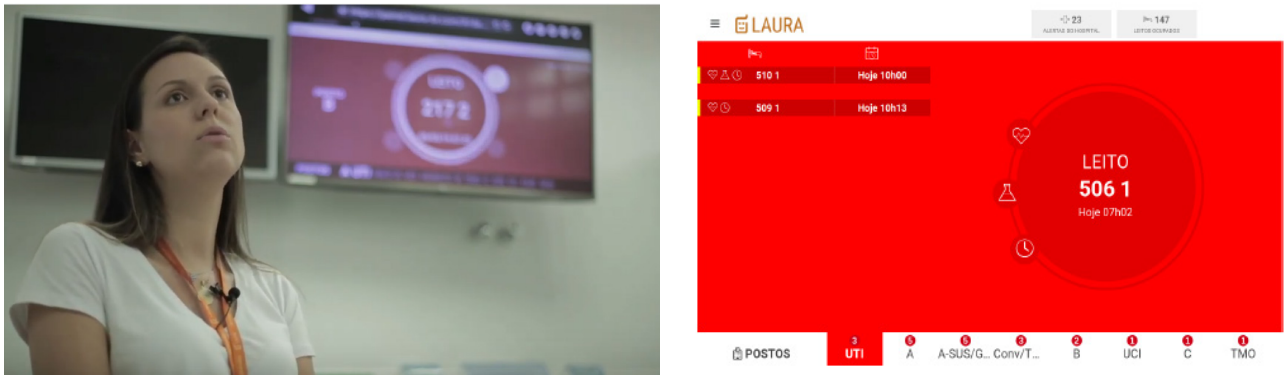
1. CÓMO FUNCIONA EL SISTEMA

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 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**:

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

  - 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;
  - 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 ($p = 0.02$).

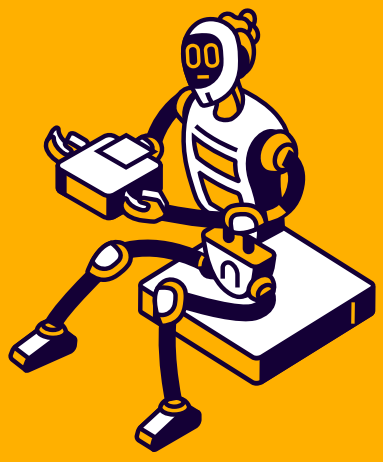
La investigación de Gonçalves et al. (2020) abordó la implantación de Laura en los aspectos vinculados a la **interacción entre el personal de enfermería y tecnología** en un hospital filantrópico durante el año 2018. Mediante observación participante y entrevistas con actores

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

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

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).

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.

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 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

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.

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'.

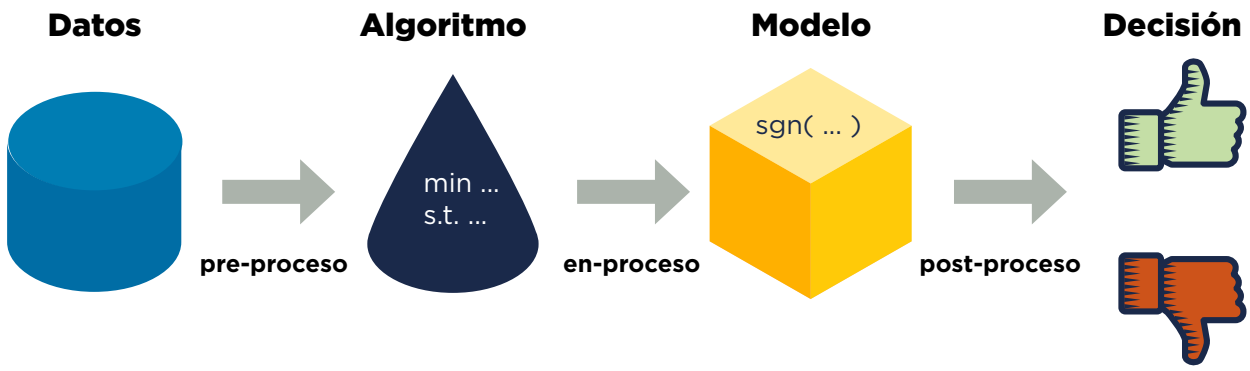
La discriminación estadística es una **discriminación grupal basada en un hecho que es estadísticamente relevante**. Un ejemplo clásico de discriminación estadística es no contratar a una mujer que reúne las competencias para un puesto laboral porque las mujeres tienen mayor probabilidad de tomar una 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).

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

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

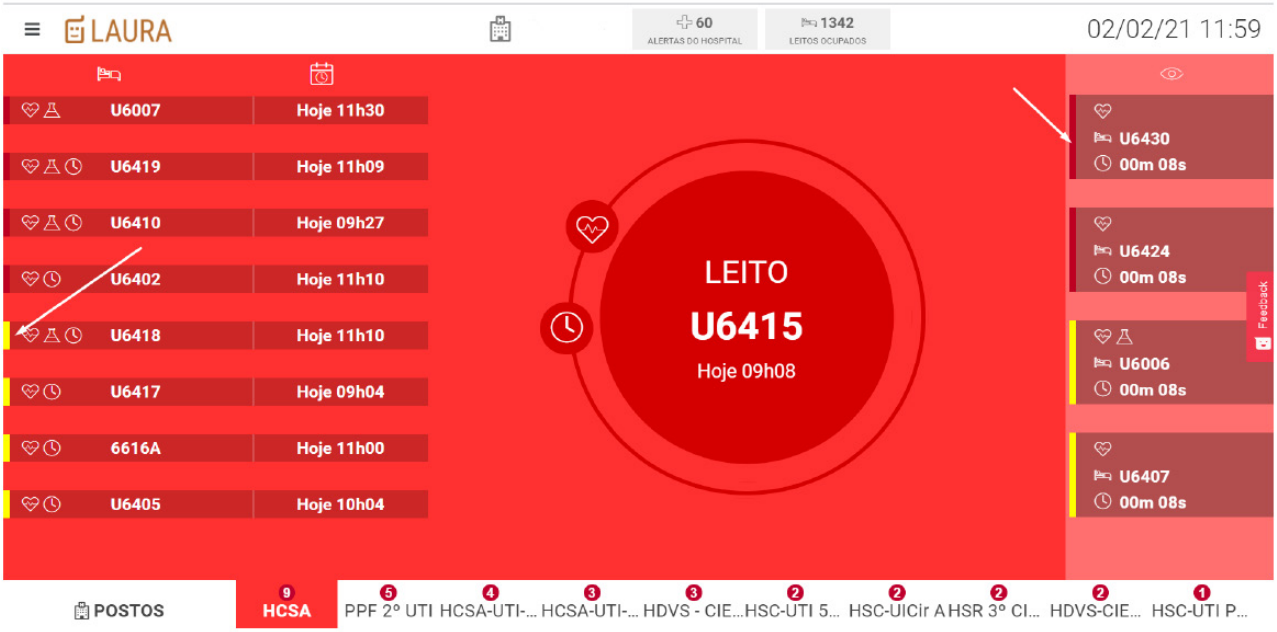
Fuente: Laura.

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

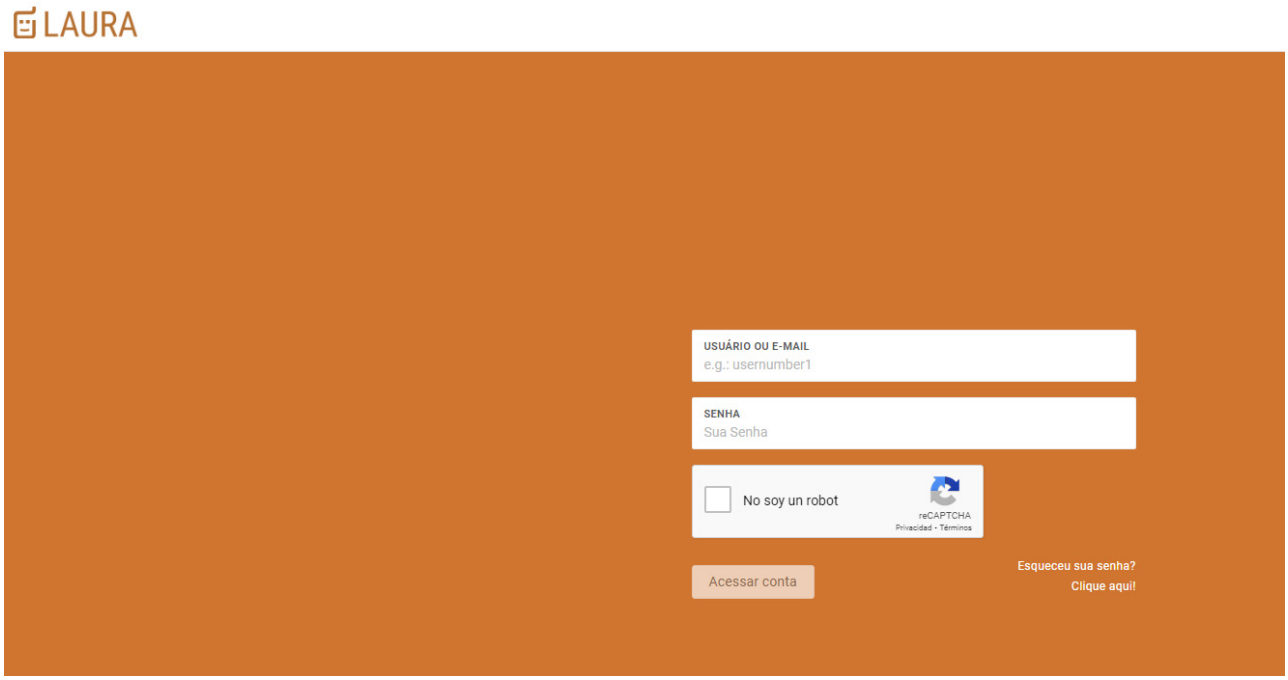
Dimensión	Análisis	Recomendaciones
Usabilidad	Se advierte claridad y coherencia en términos de inteligibilidad de las interfaces, el diseño del soporte de la información y la iconografía de los sistemas.	Se recomienda realizar encuestas interhospitalarias para establecer las limitaciones en relación 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 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.

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

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

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 %
80+	202	7.0 %

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

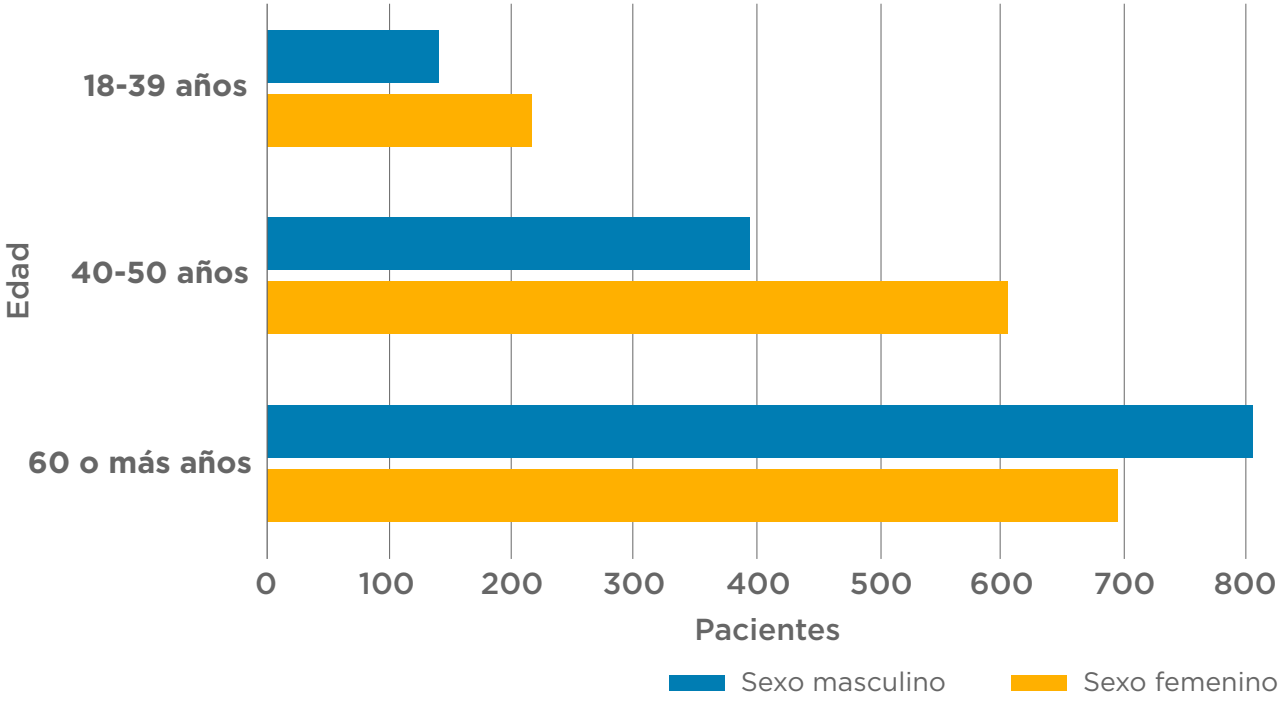
Resultado	#	%
Alta hospitalaria	2453	85.4 %
Deceso	421	14.6 %

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

Edad	**Sexo**	**Verdaderos Positivos**	**Flasos Positivos**	**PPV**	**PPV Disparity**
18-39 años	Masculino	19	9	67.9 %	**1.35**
18-39 años	Femenino	11	17	39.3 %	**0.78**
40-59 años	Masculino	58	44	56.9 %	1.13
40-59 años	Femenino	48	66	42.1 %	0.84
60 o más	Masculino	146	144	50.3 %	1.00
60 o más	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

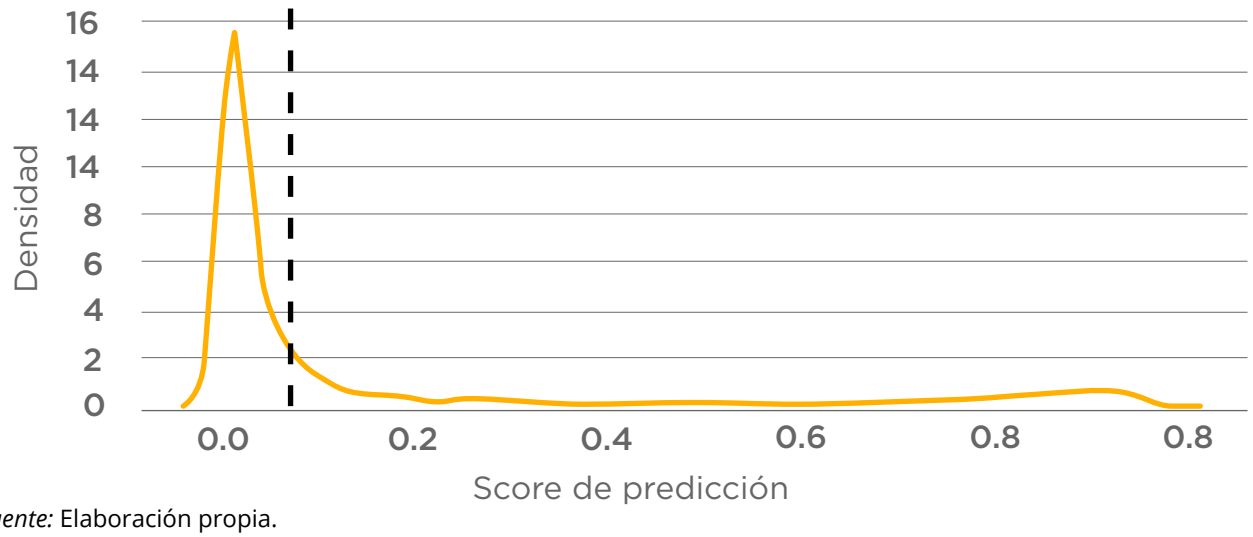
Fuente: elaboración propia.

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

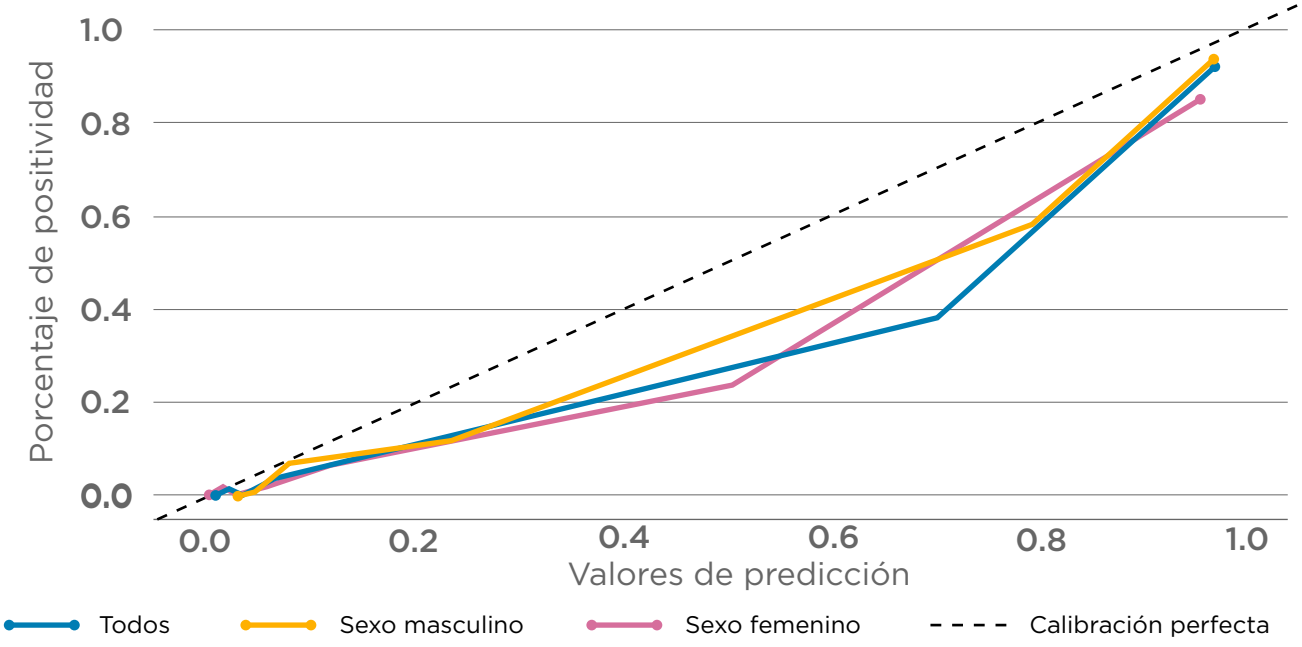
Una vez definida la curva de densidad de probabilidad, se realiza un **análisis de las curvas de calibración**, lo cual permite analizar la probabilidad de deceso en diferentes grupos. En particular, la curva de calibración indica:

- 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.

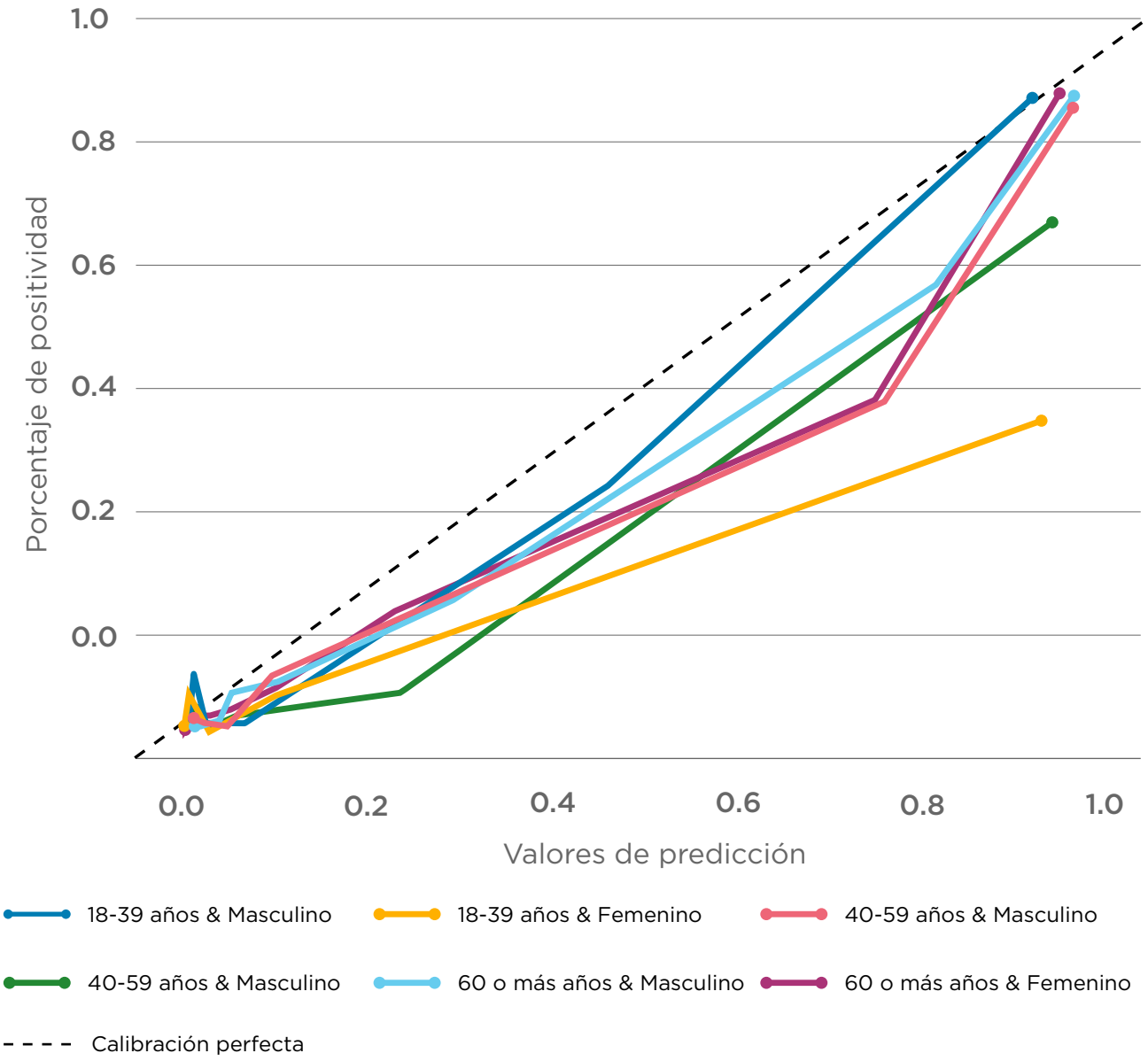
Figura 12. Curva de calibración por sexo



Fuente: Elaboración propia.

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



Fuente: Elaboración propia.

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

Dimensión	Métrica	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? *Significan*ce, 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

eticas

BID
Mejorando vidas