FAMEWS: a Fairness Auditing tool for Medical Early-Warning Systems

Marine Hoche* MARINE.HOCHE@ALUMNI.ETHZ.CH
ETH Zürich, Switzerland
Olga Mineeva* OMINEEVA@ETHZ.CH
ETH Zürich, Switzerland
MPI for Intelligent Systems Tübingen, Germany
Manuel Burger MANUEL.BURGER@INF.ETHZ.CH
ETH Zürich, Switzerland
Alessandro Blasimme ALESSANDRO.BLASIMME@HEST.ETHZ.CH
ETH Zürich, Switzerland
Gunnar Rätsch RAETSCH@INF.ETHZ.CH
ETH Zürich, Switzerland

Abstract

Machine learning applications hold promise to aid clinicians in a wide range of clinical tasks, from diagnosis to prognosis, treatment, and patient monitoring. These potential applications are accompanied by a surge of ethical concerns surrounding the use of Machine Learning (ML) models in healthcare, especially regarding fairness and non-discrimination. While there is an increasing number of regulatory policies to ensure the ethical and safe integration of such systems, the translation from policies to practices remains an open challenge. Algorithmic frameworks, aiming to bridge this gap, should be tailored to the application to enable the translation from fundamental human-right principles into accurate statistical analysis, capturing the inherent complexity and risks associated with the system. In this work, we propose a set of fairness impartial checks especially adapted to ML early-warning systems in the medical context, comprising on top of standard fairness metrics, an analysis of clinical outcomes, and a screening of potential sources of bias in the pipeline. Our analysis is further fortified by the inclusion of event-based and prevalence-corrected metrics, as well as statistical tests to measure biases. Additionally, we emphasize the importance of considering subgroups beyond the conventional demographic attributes. Finally, to facilitate operationalization, we present an open-source tool FAMEWS to generate comprehensive fairness reports. These reports address the diverse needs and interests of the stakeholders involved in integrating ML into medical practice. The use of FAMEWS has the potential to reveal critical insights that might otherwise remain obscured. This can lead to improved model design, which in turn may translate into enhanced health outcomes.

Data and Code Availability In this study, we primarily experiment with HIRID dataset (Faltys et al., 2021), which is publicly available for download on PhysioNet (Goldberger et al., 2000), and with the benchmark models for early-detection of organ failure developed by Yèche et al. (2021) whose code base is available at https://github.com/ratschlab/HIRID-ICU-Benchmark/.
The FAMEWS open-source tool is available at: https://github.com/ratschlab/famews.

Institutional Review Board (IRB) The institutional review board (IRB) of the Canton of Bern approved the study on retrospective ICU (BASEC 2016 01463). The need for obtaining informed patient consent for patient data from our institution was waived owing to the retrospective and observational nature of the study.

Authors' contributions
M.H. conceptualized the study, defined the methodology, developed the tool, conducted all the computational experiments, interpreted the results and prepared the manuscript.
O.M. conceptualized the study, defined the method-

* These authors contributed equally

ology, assisted in interpreting the results, reviewed the tool's code base, prepared the manuscript, provided supervision and feedback and coordinated the project.
M.B. assisted in data preprocessing, provided the technical support, created the Python package and reviewed the manuscript.
A.B. provided supervision and feedback and reviewed the manuscript.
G.R. conceptualized the study, secured funding, provided supervision, provided technical and conceptual feedback, and resources, reviewed the manuscript.

1. Introduction

We are witnessing the rise of Machine Learning (ML) models targeting the healthcare domain. The increasing availability of electronic health record (EHR) datasets enables the development of AI-based monitoring systems in the hospital. For instance, Yèche et al. (2021) propose benchmark models for early detection of organ failure based on the HiRID dataset (Faltys et al., 2021). These prognosis early-warning systems aim to raise the alarm in case of a high risk of organ failure within the next 12 hours. These systems are meant to be applied to critically ill patients and could have a tremendous impact on their health outcomes. As with every ML model, these systems can be biased (Coeckelbergh, 2020) and could lead to unfair health disadvantages for some patient groups (Vayena et al., 2018). Governments worldwide have expressed concern about the ethics and safe integration of ML systems. For instance, the proposed EU AI Act[1][2] aims to answer to the urgency of framing the models with strict regulatory policies. Regarding the fairness of such models, the draft of the act promotes audits of algorithms and datasets to ensure non-discrimination and non-violation of human rights. To this end, they require developers to provide documentation about the model's general characteristics, capabilities, and limitations. However, no further details are provided on how to audit fairness in practice. As highlighted in the review of algorithmic fairness (Pagano et al., 2023), this task is challenging as there is no consensus on how to measure the fairness of an algorithm.

1. https://europarl.europa.eu/doceo/document/TA-9-2022-0140_EN.html
2. https://data.consilium.europa.eu/doc/document/ST-15698-2022-INIT/EN/pdf

To fully comprehend the issue of bias in medical ML, we conducted exploratory work with ethics professionals and clinicians analyzing early detection of circulatory failure as developed in the HiRID benchmark (Yèche et al., 2021). In this first attempt (to the best of our knowledge) to design a fairness auditing framework for early-warning systems, we acknowledge the necessity to not only check for classical notions of fairness but also to investigate the fairness of the early-warning system's real-world consequences (McCradden et al., 2020). We question various system's design choices from a fairness perspective as bias can be introduced at many stages of the Machine Learning pipeline (Rajkomar et al., 2018). We summarize our learnings in an open-source tool FAMEWS which primarily complements the HiRID benchmarks (Yèche et al., 2021), but is applicable to a wide range of early-warning systems.
Our main contributions are:

1. **A flexible fairness-auditing framework tailored for clinical early-warning systems.** The framework is depicted in Figure 1. In the clinical context, patient grouping based on medical attributes such as admission type, comorbidities, or patient consciousness helps to spot model biases and identify disadvantaged subgroups beyond static demographic attributes (like race or gender). We propose grouping definitions for the HiRID dataset, but the user may change and augment them (Figure 1A). The tool is not restricted to any specific dataset, model type, or prediction task. If lacking some inputs, the user can run only part of the analysis (Figure 1B).

2. **Evaluating ML models, not only through standard metrics but also through comparison of clinical outcomes and screening of the potential sources of bias.** Available analyses are listed in Figure 1C and described in Section 3. We focus on prognosis models estimating future risk and providing early alarms, differing from classification setup by including a time dimension. Differences in timing lead to unfair outcomes as well as discrepancies in alarm's accuracy. Also, as to capture an event, it is enough to have only one alarm, we need to measure recall from the event point of view (in addition to a conventional timestep-based recall). Medical variables serve as input signals and define prediction targets. Differences in their levels and missingness patterns, even if initially clini-

cally justified, can mislead model selection and obscure fairness measurements. Differences in feature ranking across cohorts can also result in an unrepresentative model, especially while implementing a submodel reduced to the most important features. We address these concerns with the screening stages in the framework.

3. **Proposing the automatic generation of a PDF report that is easily shareable with various stakeholders and comprises the detailed fairness analysis and insightful summaries of each audit stage.** Provisioned stakeholder's needs and interests are described in Section 4.2. We don't differentiate between users while generating the report. By including all levels of analysis detail, we aim to ease communication as every stakeholder is viewing the same version of the report, and in addition, we do not hide any potentially critical information. A link to an example of the produced report is given in Appendix D, and the insights derived from it are in Appendix C.

2. Related work

In recent years, with the rise of concern surrounding the fairness of Machine Learning algorithms, tools to detect bias in these models have emerged (Bellamy et al., 2018; Weerts et al., 2023; Cabrera et al., 2019; Wexler et al., 2019; Saleiro et al., 2018; Hertweck et al., 2023). In Table 1, we summarise the characteristics of popular fairness auditing tools and compare them to our framework.

Previous works focus on fair decision-making and as such support binary classifiers. Nonetheless, some of these tools extend to multiclass classifiers or regressors, as shown in the first row of Table 1.

Group fairness can be described as the absence of systematic disadvantages towards a group of individuals that share a common attribute. The type of supported grouping is an important tool characteristic that we outline in Table 1. In the algorithmic fairness literature, classical groupings are based on protected features such as ethnicity, gender, or age and there exists a notion of a privileged and an unprivileged category. We follow the most recent tools and expand this precept by letting the user define their own grouping, which can be multicategorical. FairVis (Cabrera et al., 2019) even proposes to scan the set of possible features to find the most discriminated intersectional group.

In order to assess the fairness of a model, the Machine Learning community relies on formalizations of fairness (Makhlouf et al., 2021). They can be defined as a mathematical condition on the individual's attributes and the model output, that when satisfied ensures the model's compliance with a certain vision of fairness. To approximate these formalizations, fairness auditing tools propose to compare common performance metrics from one group to another.

The detection of unfair model outputs opens the question of where the bias is coming from. The source of bias screening is another comparison characteristic in Table 1. In Meng et al. (2022), authors explore how interpretability techniques can be used to grasp the underlying mechanics of detected biases in an ML model. For the same purpose, What-if Tool (Wexler et al., 2019) offers an interactive platform to explore trained models. For instance, they support counterfactual analysis to investigate which attributes have an unjustified effect on the prediction. While What-if tool offers a lot of capabilities to examine the model's robustness (exploration of feature importance, data distribution, and missingness), it lacks the possibility to perform these analyses per subgroup. Moreover, the counterfactual analysis on protected attributes is quite intricate to perform for medical applications as some of these attributes (such as age and sex) have direct clinically justified impacts on the label.

Finally, a couple of frameworks like FairnessLab (Hertweck et al., 2023) and Aequitas (Saleiro et al., 2018) go beyond the classical bias analysis tools by providing a more comprehensive fairness assessment. They output an intuitive summary with explanations related to relevant ethics and justice concepts, in this way becoming usable by developers as well as regulators and guiding the users to the most adequate fairness metric. For instance, the Aequitas framework (Saleiro et al., 2018) presents an interesting solution for generating fairness reports. It outputs detailed plots to compare different formalizations of fairness across groups as well as summary assessment to easily comprehend for which groups and metrics the model is biased. However, this framework is only suitable for classical binary classification, lacking event-based metrics which are key for evaluating early-warning systems. They also don't propose outcome-based metrics or screening of potential sources of bias. Moreover, the details about the statistical methodology of their work are miss-

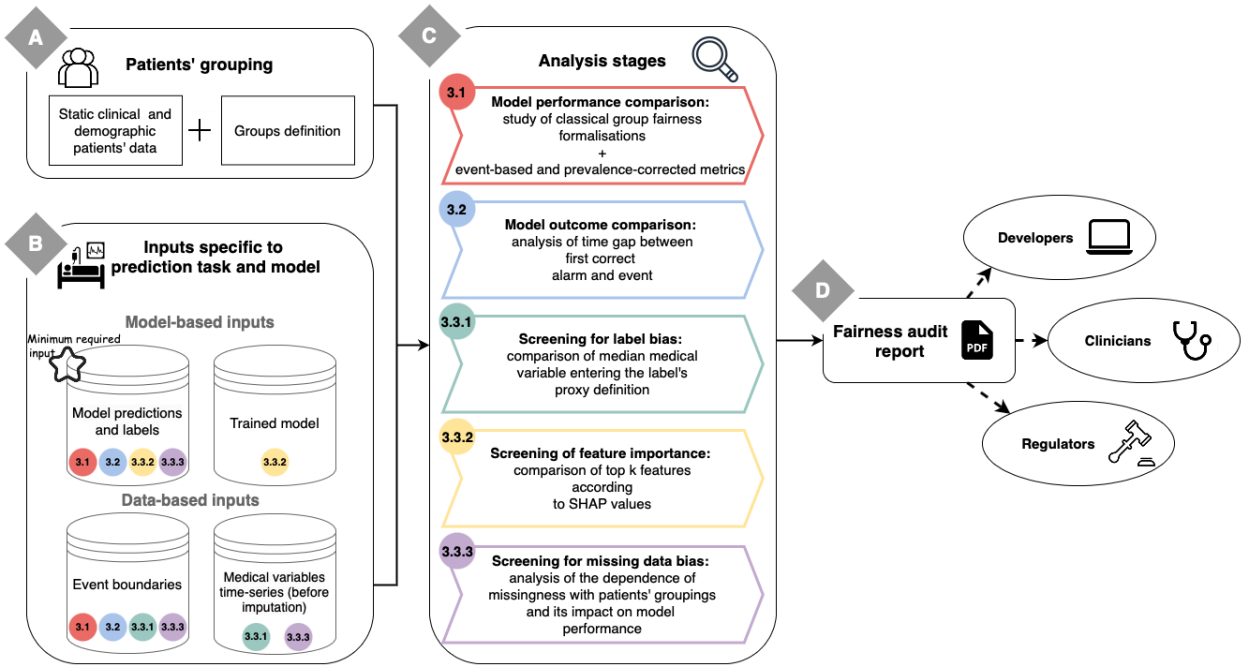

Figure 1: Schema summarising FAMEWS workflow. The user first needs to provide the patients' groupings (A), which can be based on demographics (like gender or race) or static clinical attributes (like admission reason). Then, for each prediction task and model, the user has to provide model and data-based inputs that are specific to the ML system to audit (B). Afterwards, the different analytical stages can be run (C). Their numbering indicates the corresponding section in the paper. Each analysis stage requires a specific set of inputs depicted in block B by its numbered colored dot. The results of the analyses are gathered in a PDF report that can be shared with the different stakeholders (D).

ing.

Discussed frameworks are available as libraries and some (Table 1) also embed convenient automatic visualization functionalities like a dashboard or report generation.

Focus on the medical context and early-warning systems differentiate our work from others, that are more general, but missing some essential details for this particular application.

3. Tool description

FAMEWS aims to facilitate systematic fairness audits of ML-based alarm systems in the medical field. We designed our tool to widen the usual fairness auditing scope: we assess classical fairness metrics but we also examine the fairness of clinical outcomes and investigate the potential sources of bias. Its main functionalities are summarized in Figure 1.

We consider alarm systems that take as input time-series of medical variables (lab measurements, medications, etc.) and return for each time step a score indicating how likely is the patient to undergo an event within the next X hours.

Our audit is based on comparing key statistics across cohorts of patients. The cohorts can be formed with usual demographics and static clinical information (in Figure 1A). For instance, for the HiRID dataset, the framework includes clinically relevant groupings, such as admission reasons (like trauma or cardiovascular). In the generated PDF report, we display the cohorts' composition (total number of patients and number of patients undergoing an event). We also give the possibility for the users to filter out cohorts that don't

Table 1: Comparison of fairness auditing tools

Characteristic	AI Fairness 360	Fairlearn	FairVis	What-if Tool	Aequitas	FairnessLab	**Our tool**
Other task than binary classification	✓	✓	✓	✓	✗	✗	✓
Flexible grouping (Not only binary)	✓	✓	✓	✓	✓	✗	✓
Classical fairness metrics	✓	✓	✓	✓	✓	✓	✓
Source of bias screening	✗	✗	✓	✓	✗	✗	✓
Comprehensive fairness assessment	✗	✗	✓	✗	✓	✓	✓
Robust statistical analysis	✓	✗	✗	✗	✗	✗	✓
Visual interface	✗	✓	✓	✓	✓	✓	✓

have enough patients with events (by default this parameter is set to 1), as the analysis would not be statistically significant for them.

An overview of the required inputs for each of the stages is indicated in Figure 1B by a colored dot with a section number. The minimum required input is the model's predictions and true labels for each timestep. Additionally, the time boundaries of the target events extend the audit to the assessment of performance metrics from the event scope and alarm timing comparison. Access to the trained model (or directly SHAP feature importance values) and the time series dataset allows FAMEWS to run screenings of potential sources of bias.

We recommend providing predictions from models trained with different random seeds, as this will reduce the impact of model randomness on audit results. For each stage of our audit pipeline (in Figure 1C), we run a detailed statistical analysis, that conforms to best practices, and we generate aggregated views to summarize key takeaways. These elements are gathered in a PDF report (Figure 1D). In the following paragraphs, we present the goal and motivation of each analysis stage, the metrics and statistical techniques used to capture disparities between cohorts, and outline generated visualizations and aggregated views for the fairness report.

3.1. Classical formalizations of fairness: comparison of model performance across cohorts

Goal In this stage, we compare the model's performance and the validity of the threshold choice across different patient cohorts through classical fairness notions (Makhlouf et al., 2021; Chen et al., 2023). An example can be found in section 2 of the sample report (Appendix D).

Metrics For each cohort of patients, we compute the metrics related to a set of adequate fairness notions (they are listed in Appendix A together with the definitions of the performance metrics, for readers new to the field we highly recommend exploring some online tutorials[3][4] on the topic). We implemented binary (recall, precision, FPR, and NPV) and score-based metrics (AUROC, AUPRC, average score on positive and negative classes, calibration error) as they are relevant at different phases of model development. For instance, while tuning the model, score-based metrics are valuable, whereas a deployed model with binary outputs is evaluated using relevant binary metrics. For our targeted medical application, it is beneficial to consider event-based metrics such as event-based recall (number of predicted events over the total number of events) and event-based AUPRC (area under the precision / event-based recall curve). We added the possibility of comparing the precision, NPV, and AUPRC after correction for prevalence (due to the imbalance of positive labels across cohorts). This is equivalent to comparing the original version of these metrics assuming the cohorts have equal prevalence (more details are given in Appendix B).

3. https://developers.google.com/machine-learning/crash-course/fairness
4. https://developers.google.com/machine-learning/glossary/fairness

Statistical methodology We compare the metrics for each cohort to the rest of the patients. To ensure the statistical robustness of this comparison, we first bootstrap the patient population of the test set (we draw with replacement 100 random samples of the test set size) and compute for each cohort in each sample the metrics listed above. We then perform the Mann-Whitney U test with Bonferroni correction. From these statistical tests we obtain for each metric the categories of patients which are significantly worse off compared to the rest of the population. We then quantify these disparities by computing the absolute difference in median metric (taken over the bootstrapped samples) between patients of the category and patients outside it:

$$\Delta = \Big| median\left\{metric_{p\in S_n\cap G}\right\}_{n=1}^{N} - median\left\{metric_{p\in S_n\cap \bar{G}}\right\}_{n=1}^{N}\Big| \quad (1)$$

with S_n the n^{th} bootstrapped sample, N the total number of bootstrapped samples drawn, G the studied cohort and $\bar{G}$ the rest of patients.

Visualizations The results of the comparison are presented as tables in the report. We display box plots for each metric with the median, first quartile, and third quartile over the bootstrapped samples. Cohorts that are significantly worse off are highlighted with a star (see sample report Figure 2.1.1.a p.9). For the score-based metrics, we report performance curves: calibration, ROC, and precision-recall (also event-based) curves (see sample report Figure 2.1.1.b p.12). The colored error area represents the standard deviation computed over the bootstrap samples. To ease comparison, we keep the same scale for each metric across the entire report.

Aggregated views 3 aggregated views are proposed for this stage:

1. Summary statistics for each metric and grouping: it is composed of the macro-average, the minimum over the grouping's categories, and the metric value for the minority category (see sample report 2.1.1 p.5).

2. Summary view based on the ratio of significantly worse metrics: For each cohort, we report the ratio of significantly worse metrics over the total number of analyzed metrics. We highlight which category of patients within the grouping and across all groupings is the worst in terms of ratio. The largest delta, as defined in Equation (1), for this category is stated (see sample report 2.1.2 p.7).

3. Table displaying for each metric the 3 cohorts with the largest delta that are significantly worse off than the rest of the population. They are also flagged with a red star on the corresponding metric box plot (see sample report 2.1.3 p.8).

3.2. Checking for bias of outcomes: comparison of the time gap between first correct alarm and event across cohorts

Goal One outcome of the early-warning system is to direct additional clinical attention to specific patients to prevent the forecasted events. We analyze whether the alarm is triggered sufficiently in advance for the different cohorts of patients. An example is in section 3 of the sample report (Appendix D).

Metrics For each detected event, we compute the time gap between the first correct alarm and the event. The bigger the time gap the better off a patient is.
To not bias this analysis, we first split the events with respect to how much time in advance the alarm could be triggered. For the sake of clarity, let us consider an alarm system with a 12-hour horizon. If an event happens three hours after the start of the stay, the alarm can be triggered at most 3 hours in advance; while if it occurs after 24 hours, the alarm can be raised 12 hours in advance. It is thus not equitable to compare these two categories of events. To overcome this issue, we propose to split the possible alarm window into 4 (configurable) parts: 0-3h, 3-6h, 6-12h, and more than 12h. For each of our alarm window splits and cohort of patients, we then compute the median time gap.

Statistical methodology We draw 100 bootstrap samples (as for the previous stage in Section 3.1). For each bootstrapped sample, each alarm window split, and each cohort of patients, we compute the median time gap. We then use the Mann-Whitney U test with Bonferroni correction to determine which cohorts are significantly worse off than the rest of the population. We quantify the disparity by computing the difference between the median (taken over the bootstrapped samples) time gap for patients belonging to a cohort and patients not belonging to it, for each window split. This is equivalent to computing Δ in Equation (1) with *metric* being the median time

gap for the events falling into a specific window split for a selected cohort.

Visualizations The comparison results are outlined in tables and visually displayed in box plots, in the same fashion as for our first analysis (Section 3.1)(see sample report Figure 3.2.1.a p.35).

Aggregated views 2 aggregated views are proposed for this stage:

1. Summary statistics for each alarm window split and grouping of patients composed of the macro-average,the minimum metric value over all the grouping's categories, and the value for the minority category (see sample report 3.1.1 p.33).

2. Table displaying for each alarm window split the 3 cohorts with the biggest delta that are significantly worse-off than the rest of the population. These cohorts are also flagged with a red star on the corresponding box plot (see sample report 3.1.2 p.34).

3.3. Assessing level of bias for potential sources

3.3.1. Comparison of some medical variables across cohorts

Goal It is quite common in clinical contexts to rely on proxy labels instead of ground truth to depict a medical phenomenon. For instance, circulatory failure can be defined through arterial lactate and blood pressure levels. This analysis has been specially designed to tackle the problem of label bias (Wick et al., 2019; Rateike et al., 2022) that can occur in these settings. We want to check whether the proxy used to define the label is correct for all cohorts. An ill-defined label can create degradation in performance and unfair outcomes. We thus propose to compare the distribution of medical variables used in the proxy definition across the different cohorts of patients. Nonetheless, this stage can also be used to study other time-series variables that are relevant to the user. An example can be found in section 4 of the sample report (Appendix D).

Metrics For each cohort, we compare the distribution of chosen medical variables to the rest of the population. For each patient, we compute the median value over the entire stay. According to this stage's goal, we expect that undergoing an event has a strong influence on the variable value. We thus also inspect separately periods of stay free of events and patients without events.

Statistical methodology We draw 100 bootstrap samples from the train set in the same fashion as in Section 3.1. For each sample and each cohort, we end up with three different median values (for all data points, not during events, and for patients free of events) for the selected medical variables. We compare the distribution of each median from one cohort to the rest of the population using the Mann-Whitney U test with Bonferroni correction. We quantify the difference in median values by computing the absolute difference in medians (median taken over the bootstrapped samples of the different medians) between patients belonging to a cohort and patients not belonging to it. This is equivalent to computing Δ in Equation (1) with *metric* being one of the three median values for a medical variable and a selected cohort.

Visualizations We report the results in tables and with box plots (see sample report Figure and Table 4.2.1.a p.42). The star on these plots flags the categories of patients with a significantly different median value compared to the rest of the patients.

Aggregated views We outline, for each of the selected medical variables and the median computation methods, the 3 cohorts with the biggest delta in median value that are significantly different from the rest of the population (see sample report 4.1.1 p.41). These cohorts are also signaled with a red star on the corresponding variable box plot.

3.3.2. Comparing the top k features across cohorts

Goal Regarding explainability concerns, it is essential for the stakeholders to know the features that drive the prediction process. We check whether feature importance deviates across patient cohorts. We consider this to be of special interest for two scenarios. First, while considering a submodel developers usually keep only the most important features from the validation set (Hyland et al., 2020), however in this process, they can disregard features that are important to minority cohorts, losing predictive power for them (Zong et al., 2023). Then, to check the clinical relevance of the model, it can be useful to show medical practitioners, not only the global top features but also the top features for the different subcohorts. Indeed they might want to review how the medical

variables impact the model prediction depending on the various patient profiles. An example is in section 5 of the sample report (Appendix D).

Metrics To study the feature importance, we will rely on SHAP values (Lundberg and Lee, 2017). This is a local explanation method, allowing us to obtain the feature importance for each data point. We can thus obtain the feature importance for each patient and aggregate them per cohort. Furthermore, this method aligns better with human intuition than other feature importance estimation techniques (Lundberg and Lee, 2017), such as LIME (Ribeiro et al., 2016). Nonetheless, this framework can yield inaccurate feature importance values when features are dependent or correlated. (Aas et al., 2021).
For each patient and a given feature, we thus quantify its importance with the absolute mean SHAP value over the stay. Then we derive a feature ranking for a cohort based on the mean feature importance over all of its patients. We compare the feature ranking of each cohort to the global feature ranking using a similarity measure on lists called the rank-biased overlap (RBO) (Webber et al., 2010). This measure has the particularity of giving more weight to the head compared to the tail (weighting parameter $p = 0.935$). This aspect is particularly suitable to the comparison of feature importance rankings as we care more about differences for the top features (Sarica et al., 2022). Nonetheless, this property highly depends on the weighting parameter, which can be challenging to tweak properly. For each feature ranking, we flag the features that significantly changed rank compared to the global ranking.

Statistical methodology To establish the statistical relevance of our analysis, we compute the RBO for feature ranking on random simulated patient cohorts. This yields an upper bound,

$$\min \bigcup_{i=1}^{100} \{RBO\,(rk_g, rk_{all})\}_{g \in G^i_{random}}$$

(with G^i the i^{th} random grouping, rk_g the ranking obtained on one cohort of G^i and rk_{all} the overall ranking) below which the RBO testifies of significantly different feature rankings. From these random groupings, we compute for each feature, the delta of inverse rank $\left|\frac{1}{k_{all}} - \frac{1}{k_0}\right|$ (with k_{all} the global rank and k_0 the rank we want to compare to) and obtain a lower bound,

$$\max \bigcup_{i=1}^{100} \left\{ \left| \frac{1}{k_g} - \frac{1}{k_{all}} \right| \right\}_{g \in G^i_{random}}$$

(with G^i the i^{th} random grouping, k_g the rank of the studied feature for one cohort of G^i and k_{all} its global rank) above which the delta of inverse rank indicates that the feature has a significantly different rank compared to the global ranking.

Visualizations For each cohort, we outline the top k features, we print the feature name in red when it isn't part of the global top k ranking and in blue when it changes rank within the top k ranking from global to cohort-based (see sample report Table 5.2.1.a p.51). We only color the names when the change of rank is significant. However, for each feature that changes rank, we put in parenthesis the difference in rank and the direction of change.

Aggregated views We display the RBO for each cohort, colored in red when it is significantly low (see sample report 5.1.1 p.50).

3.3.3. Comparing the missingness of key medical variables and its impact across cohorts

Goal The intensity of measurement of medical variables highly depends on their nature and the health status of the patient. As such, data used for medical applications aren't missing at random. We thus investigate how the intensity of measurement for relevant variables correlates with patients' attributes. From a fairness perspective, we can wonder whether disparities in the intensity of measurement across cohorts of patients are purely motivated by medical reasons or whether some forms of discrimination are present. We thus inspect the impact of missingness on the model performance (Getzen et al., 2023). The results could hint at adapting the data collection or the imputation practices. An example can be found in section 6 of the sample report (Appendix D).

Metrics For this analysis, the user needs to provide, for each patient, the time series of medical variables resampled on a fixed time-step grid before data imputation. For each of the selected medical variables, we forward propagate the measurement value according to its usual sampling interval (that has been indicated by the user).

First, we measure the intensity of measurements I for each patient that has at least one valid value:

$$I = 1 - \frac{N_m}{N_e}$$

with N_e the number of expected measurements and N_m the number of missed measurements. N_e is defined as $N_e = \frac{los}{t_e}$ with los the patient's length of stay and t_e the expected sampling interval. N_m is obtained by summing the number of measurements that could have been done during each period $T_i^{missing}$ without valid measurements (even after propagation): $N_m = \sum_i \frac{T_i^{missing}}{t_e}$. The user provides categorization for the intensity of measurement values. For our example report, we class values below 90% as *insufficient* and above as *enough*. We put apart patients without any measurement. Then, we assess the impact of missing values on performance. The methodology is similar to the stage in Section 3.1: we measure classical metrics but instead of grouping the data points per cohort of patients, we group them based on their missingness status. Data points without valid value after propagation are grouped in the *missing_msrt* category, those belonging to patients without measurement in *no_msrt* and the rest in *with_msrt*. For this analysis, we don't measure event-based metrics. For variables used in the label's definition, it is not possible to run the analysis on the *no_msrt* category.

Statistical methodology We run the Chi-squared independence test to assess the dependence between the patients' grouping and the intensity of measurement categories.
The statistical tests for the impact of performance analysis are run in the same fashion as in Section 3.1. However, instead of comparing each cohort to the rest of the population, we compare the missingness categories *no_msrt* and *missing_msrt* against the *with_msrt* category.

Visualizations For the intensity of measurements analysis, we provide for each cohort a bar plot displaying the percentage of patients belonging to each intensity category (see sample report Figure 6.2.1.a p.56). The dotted lines show the percentage over the entire population of patients as references. For the impact on performance, we present the results in tables and box plots as in Section 3.1 (see sample report Figure and Table 6.2.2.a p.58-60).

Aggregated views For each of the selected medical variables, if the grouping and the intensity of measurements are dependent, the grouping is outlined in a table. Also, the category for the corresponding grouping with the biggest rate of patients without measurement and the one with an insufficient number of measurements are indicated. To summarize the impact on the performance, the ratio of metrics that are significantly worse than the *with_msrt* metrics is displayed for each missingness category as well as the worst delta in metrics (see sample report 6.1.1 p.55).

4. Discussion

In this paper, we described FAMEWS – a fairness auditing tool tailored for medical early-warning systems. Our approach extends the scope of classical fairness assessment tools by including an analysis of fairness of outcomes, screening of potential sources of bias, and proposing to consider clinical attributes on top of classical demographic features for fairness analysis. We will now discuss the flexibility of our tool, how our generated report can be used by the different stakeholders as well as the strengths and limitations of our work.

4.1. Flexibility of the tool

We primarily built our tool to audit the fairness of an LGBM (Light Gradient-Boosting Machine) early-warning system detecting circulatory failures in the intensive care unit on the HiRID dataset (Yèche et al., 2021). Nonetheless, we conceived it with a certain level of flexibility, allowing it to be extended to a broader range of applications. We tested our framework on other alarm systems (early detection of respiratory failure (Hüser et al., 2024)) with different alarm-to-event horizon lengths, on other datasets like MIMIC-III (Johnson et al., 2016), and on other types of models (Long Short-Term Memory networks). The users can define their own patients' groupings depending on the available attributes, provide processed inputs rather than raw data, or run only a subset of the stages if they don't have access to some input data. Moreover, some stages can be adapted to audit other types of binary classifiers; for instance, where the model outputs for each patient a single prediction instead of a time series. Finally, our tool is open-source, offering the possibility to the users to further extend its functionalities.
However, to complete this fairness audit, the user needs a minima access to some test data and the ca-

pability to generate predictions from the model (see Figure 1B).

4.2. Intended use of the produced report

We designed our report as a conveniently exchangeable document that can be understood and used by different stakeholders. We decided against an interactive dashboard that, although more convenient for exploratory work, would have required the technical skills of the end user, secured access to medical data, and would not have been easily exportable. We now list the provisioned use of the report for the identified stakeholders:

Developers

- Compare different model design choices (model type, preprocessing, feature engineering) in order to choose the best model from a fairness point of view. A quick glimpse of how the model is evolving can be obtained by comparing the aggregated views of the respective reports.
- Identify targets for bias mitigation and measure the impacts of different debiasing methods. The aggregated views can be used to facilitate model comparison and choose the best bias mitigation.
- Monitor the behavior of the model, from a fairness point of view, while using the model on new data samples or retraining it (after the deployment for instance).

Clinicians

- Adapt their reliance on the model by learning about its main biases, which are highlighted in the aggregated views. For instance, if the practitioners are aware that the model is performing worse for a specific patient cohort then they will not overly rely on the model to monitor these patients, avoiding falling into an automation bias (Rajkomar et al., 2018).
- Provide developers feedback and help them to comprehend certain disparities, especially in the screening of sources of bias analyses. For instance, the results of the label bias screening can be used to discuss the validity of the label proxy definition for all patients. Their feedback can then guide the developers in choosing adequate bias mitigation techniques.

Regulators

- Get informed about the model limitations in terms of bias and obtain a brief overview of the demographics.
- Check that the model complies with actual regulations in terms of fairness and non-discrimination.

4.3. Strengths and limitations of our framework

The resulting audit report might seem cumbersome to apprehend. We nonetheless believe it is necessary to present the entire analysis in the report, as selecting relevant results is subjective and might hide relevant disparities to the end users. We facilitate its navigation with a table of contents, a glossary, and aggregated views for each analysis stage. These views help in grasping the main takeaways of the report. However, like every summary, it is not self-sufficient and we insist on the necessity to refer to the more detailed analyses to fully understand the extent of potential biases.

Despite its size, our report is rather limited in the range of screened sources of bias. We tackle the ones that we deem crucial for our prime use case. However, depending on the system's design choices, other sources are also valuable to explore. We acknowledge similar limitations on our exploration of bias of outcomes. Indeed, this issue is deeply dependent on the application and some are not measurable without access to the actual real-world consequences of the ML system. For instance, we don't address the issue of censored data, which can occur in the context of warning systems in real-word setting in the medical domain, while it can represent a real fairness challenge (Zhang et al., 2023).We thus encourage the users to extend the fairness audit to the inspection of post-deployment biases. Then, our tool proposes a limited set of fairness metrics, contrary to other tools. Nonetheless, we implemented evaluation with event-based metrics and prevalence correction which we didn't find in other fairness auditing tools, but we consider them important for early-warning systems auditing. Finally, we enforced best statistical practices to bring an adequate level of robustness to our audit results. We realized that this aspect was missing in existing fairness analysis frameworks.

In summary, we propose FAMEWS to assess the fairness of ML-based early-warning systems. We be-

lieve that the wide adoption of such auditing tools could ease the communication between regulators, developers, and clinicians and could assist in developing both accurate and ethical applications.

References

Kjersti Aas, Martin Jullum, and Anders Løland. Explaining individual predictions when features are dependent: More accurate approximations to shapley values. *Artificial Intelligence*, 2021.

Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John T. Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang. AI fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *CoRR*, 2018.

Ángel Alexander Cabrera, Will Epperson, Fred Hohman, Minsuk Kahng, Jamie Morgenstern, and Duen Horng Chau. Fairvis: Visual analytics for discovering intersectional bias in machine learning. *CoRR*, 2019.

Richard Chen, Judy Wang, Drew Williamson, Tiffany Chen, Jana Lipkova, Ming Lu, Sharifa Sahai, and Faisal Mahmood. Algorithmic fairness in artificial intelligence for medicine and healthcare. *Nature biomedical engineering*, 2023.

Mark Coeckelbergh. In *AI Ethics*, chapter Bias and the Meaning of Life. The MIT Press, 2020.

Martin Faltys, M. Zimmermann, X. Lyu, Matthias Hüser, S. Hyland, Gunnar Rätsch, and T. Merz. Hirid, a high time-resolution icu dataset (version 1.1.1). *PhysioNet*, 2021.

Emily Getzen, Lyle Ungar, Danielle Mowery, Xiaoqian Jiang, and Qi Long. Mining for equitable health: Assessing the impact of missing data in electronic health records. *Journal of Biomedical Informatics*, 2023.

A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. Ch. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation*, 2000.

Corinna Hertweck, Joachim Baumann, Michele Loi, Eleonora Viganò, and Christoph Heitz. A justice-based framework for the analysis of algorithmic fairness-utility trade-offs, 2023.

Matthias Hüser, Xinrui Lyu, Martin Faltys, Alizée Pace, Marine Hoche, Stephanie L. Hyland, Hugo Yèche, Manuel Burger, Tobias M. Merz, and Gunnar Rätsch. A comprehensive ml-based respiratory monitoring system for physiological monitoring & resource planning in the icu. *medRxiv*, 2024.

Stephanie L Hyland, Martin Faltys, Matthias Hüser, Xinrui Lyu, Thomas Gumbsch, Cristóbal Esteban, Christian Bock, Max Horn, Michael Moor, Bastian Rieck, et al. Early prediction of circulatory failure in the intensive care unit using machine learning. *Nature medicine*, 26(3):364–373, 2020.

A.E Johnson, T.J Pollard, L Shen, LW Lehman, M Feng, M Ghassemi, B Moody, P Szolovits, L.A Celi, and R.G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 2016.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17. Curran Associates Inc., 2017.

Karima Makhlouf, Sami Zhioua, and Catuscia Palamidessi. On the applicability of machine learning fairness notions. 2021.

Melissa McCradden, Shalmali Joshi, Mjaye Mazwi, and James Anderson. Ethical limitations of algorithmic fairness solutions in health care machine learning. *The Lancet Digital Health*, 2020.

Chuizheng Meng, Loc Trinh, Nan Xu, James Enouen, and Yan Liu. Interpretability and fairness evaluation of deep learning models on mimic-iv dataset. *Scientific Reports*, 2022.

Tiago P. Pagano, Rafael B. Loureiro, Fernanda V. N. Lisboa, Rodrigo M. Peixoto, Guilherme A. S. Guimarães, Gustavo O. R. Cruz, Maira M. Araujo, Lucas L. Santos, Marco A. S. Cruz, Ewerton L. S. Oliveira, Ingrid Winkler, and Erick G. S. Nascimento. Bias and unfairness in machine learning models: A systematic review on datasets, tools,

fairness metrics, and identification and mitigation methods. *Big Data and Cognitive Computing*, 2023.

Alvin Rajkomar, Michaela Hardt, Michael D. Howell, Greg S. Corrado, and Marshall H. Chin. Ensuring fairness in machine learning to advance health equity. *Annals of Internal Medicine*, 2018.

Miriam Rateike, Ayan Majumdar, Olga Mineeva, Krishna P. Gummadi, and Isabel Valera. Don't throw it away! the utility of unlabeled data in fair decision making. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22. Association for Computing Machinery, 2022.

Marco Ribeiro, Sameer Singh, and Carlos Guestrin. "why should I trust you?": Explaining the predictions of any classifier. In John DeNero, Mark Finlayson, and Sravana Reddy, editors, *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*. Association for Computational Linguistics, 2016.

Pedro Saleiro, Benedict Kuester, Abby Stevens, Ari Anisfeld, Loren Hinkson, Jesse London, and Rayid Ghani. Aequitas: A bias and fairness audit toolkit. *CoRR*, 2018.

Alessia Sarica, Andrea Quattrone, and Aldo Quattrone. Introducing the rank-biased overlap as similarity measure for feature importance in explainable machine learning: A case study on parkinson's disease. *Brain Informatics*, 2022.

Effy Vayena, Alessandro Blasimme, and I. Glenn Cohen. Machine learning in medicine: Addressing ethical challenges. *PLOS Medicine*, 2018.

William Webber, Alistair Moffat, and Justin Zobel. A similarity measure for indefinite rankings. *ACM Trans. Inf. Syst.*, 2010.

Hilde Weerts, Miroslav Dudík, Richard Edgar, Adrin Jalali, Roman Lutz, and Michael Madaio. Fairlearn: Assessing and improving fairness of ai systems, 2023.

James Wexler, Mahima Pushkarna, Tolga Bolukbasi, Martin Wattenberg, Fernanda B. Viégas, and Jimbo Wilson. The what-if tool: Interactive probing of machine learning models. *CoRR*, 2019.

Michael L. Wick, Swetasudha Panda, and Jean-Baptiste Tristan. Unlocking fairness: a trade-off revisited. In *Neural Information Processing Systems*, 2019.

Hugo Yèche, Rita Kuznetsova, Marc Zimmermann, Matthias Hüser, Xinrui Lyu, Martin Faltys, and Gunnar Rätsch. Hirid-icu-benchmark — a comprehensive machine learning benchmark on high-resolution icu data. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021.

Wenbin Zhang, Tina Hernandez-Boussard, and Jeremy Weiss. Censored fairness through awareness. In *AAAI Conference on Artificial Intelligence*, 2023.

Yongshuo Zong, Yongxin Yang, and Timothy Hospedales. MEDFAIR: Benchmarking fairness for medical imaging. In *The Eleventh International Conference on Learning Representations*, 2023.

Appendix A. Formalizations of fairness and performance metrics

In this section, we define in Table 2 the different performance metrics available in FAMEWS. We show in Table 3 the formalizations of fairness that we thought important to consider while auditing alarm systems in clinical settings and we link them to their corresponding performance metrics. Precision-recall curve and AUPRC aren't present in this table, as checking together for equal precision and recall across cohorts doesn't match one of the conventional notions of fairness. Nonetheless, we still include them in our audit pipeline as they are valuable performance metrics for our use-case.

Table 2: Performance metrics definitions. Definition of each of the model's performance metrics used in the first step of our fairness analysis. In the formulas, P stands for the number of positive labels, TP the number of correctly predicted positive labels, TN the number of correctly predicted negative labels, FP the number of instances with true negative labels that were incorrectly predicted as positive by the model, and FN the number of instances with true positive labels that were incorrectly predicted as negative by the model.

Performance metric	Definition
Recall	TP/P
False positive rate (FPR)	$FP/(FP+TN)$
Precision	$TP/(TP+FP)$
Negative predictive value (NPV)	$TN/(TN+FN)$
Average score on positive class	For all positive labels, average of the output scores
Average score on negative class	For all negative labels, average of the output scores
Calibration curve	The frequency of positive labels vs the mean predicted scores, it illustrates how well the probabilistic predictions of the model are calibrated
Calibration error	Area between the calibration curve and the perfect calibration line
Receiver operating characteristic (ROC) curve	True positive rate vs False positive rate
AUROC	Area under the ROC curve
Precision-recall curve	Precision vs Recall
AUPRC	Area under the precision-recall curve

Appendix B. Proof prevalence correction

Consider $\mathcal{C}$ to be a random binary classifier. It assigns class 0 and class 1 with equal probability. Let $\mathcal{D}_1$ and $\mathcal{D}_2$ be two datasets with different prevalence pv_1 and pv_2, w.l.o.g. we assume $pv_1 < pv_2$.
This classifier being random, we expect it to have the same performance on $\mathcal{D}_1$ and $\mathcal{D}_2$. Let us express the recall, FPR, precision, and NPV on both datasets.
We denote by P_1 (resp. P_2) the number of positive labels in $\mathcal{D}_1$ (resp. $\mathcal{D}_2$), N_1 (resp. N_2) the number of negative labels in $\mathcal{D}_1$ (resp. $\mathcal{D}_2$), TP_1 (resp. TP_2) the number of correctly predicted positive labels in $\mathcal{D}_1$ (resp. $\mathcal{D}_2$), TN_1 (resp. TN_2) the number of correctly predicted negative labels in $\mathcal{D}_1$ (resp. $\mathcal{D}_2$), FP_1 (resp. FP_2) the number of negative labels wrongly predicted as positives in $\mathcal{D}_1$ (resp. $\mathcal{D}_2$) and FN_1 (resp. FN_2) the number of positive labels wrongly predicted as negatives in $\mathcal{D}_1$ (resp. $\mathcal{D}_2$).

$$recall_1 = \frac{TP_1}{P_1} = \frac{0.5 \times P_1}{P_1} = 0.5 = recall_2$$

$$FPR_1 = \frac{FP_1}{N_1} = \frac{0.5 \times N_1}{N_1} = 0.5 = FPR_2$$

$$\begin{aligned} precision_1 &= \frac{TP_1}{TP_1 + FP_1} = \frac{0.5 \times P_1}{0.5|\mathcal{D}_1|} = \frac{0.5 pv_1 |\mathcal{D}_1|}{0.5|\mathcal{D}_1|} \\ &= pv_1 \\ precision_2 &= pv_2 \end{aligned}$$

$$\begin{aligned} NPV_1 &= \frac{TN_1}{TN_1 + FN_1} = \frac{0.5 \times N_1}{0.5|\mathcal{D}_1|} \\ &= \frac{0.5(1 - pv_1)|\mathcal{D}_1|}{0.5|\mathcal{D}_1|} \\ &= 1 - pv_1 \\ NPV_2 &= 1 - pv_2 \end{aligned}$$

Recall and FPR are equal for both datasets as expected. However, this is not the case for precision and NPV. Let us find a way to modify the formula of precision and NPV such that they are equal for both datasets.

Table 3: Relation between popular formalizations of fairness and performance metrics. We selected a set of formalizations of fairness that we deemed relevant for our use-case. In this table, we outline for each formalization the corresponding metrics we inspected. We consider that a notion of fairness is respected when the corresponding metric is equal across cohorts. When we use the symbol '&' that means that both metrics have to be equal. For curves, we inspect visually whether they are similar across cohorts and use their respective error metrics to assess more precisely the disparities.

Formalisation of fairness	Related performance metric
Equality of opportunity	Recall
Predictive equality	FPR
Equalized odds	AUROC, ROC curve, recall & FPR
Predictive parity	Precision
Conditional use accuracy	NPV & precision
Balance on positive class	Average score on positive class
Balance on negative class	Average score on negative class
Calibration	Calibration curve, calibration error

Correction of precision We want $c\_precision_1 = c\_precision_2$ (with $c\_precision$ the corrected precision.) We keep the higher prevalence pv_2 as a reference and we want to correct for pv_1. We denote by s the correction factor. We will artificially modify the number of false positives for $\mathcal{D}_1$ by the factor s.

$$
\begin{aligned}
&c\_precision_1 = c\_precision_2 = precision_2 = pv_2 \\
\implies &\frac{TP_1}{TP_1 + sFP_1} = pv_2 \\
\implies &\frac{0.5pv_1 \times |\mathcal{D}_1|}{0.5pv_1 \times |\mathcal{D}_1| + s \times 0.5(1-pv_1) \times |\mathcal{D}_1|} = pv_2 \\
\implies &\frac{pv_1}{pv_1 + s(1-pv_1)} = pv_2 \\
\implies &s = \frac{pv_1 - pv_1 pv_2}{pv_2(1-pv_1)} \\
&s = \frac{\frac{1}{pv_2} - 1}{\frac{1}{pv_1} - 1}
\end{aligned}
$$

Correction of NPV We want $c\_NPV_1 = c\_NPV_2$ (with c_NPV the corrected NPV). We keep the smaller prevalence pv_1 as a reference and we want to correct for pv_2. We denote by s the correction factor. We will artificially modify the number of false negatives for $\mathcal{D}_1$ by the factor s.

$$
\begin{aligned}
&c\_NPV_2 = c\_NPV_1 = NPV_1 = 1 - pv_1 \\
\implies &\frac{TN_2}{TN_2 + sFN_2} = 1 - pv_1 \\
\implies &\frac{0.5(1-pv_2) \times |\mathcal{D}_2|}{0.5(1-pv_2)|\mathcal{D}_2| + s0.5pv_2|\mathcal{D}_2|} = 1 - pv_1 \\
\implies &\frac{1-pv_2}{1-pv_2 + spv_2} = 1 - pv_1 \\
\implies &s = \frac{pv_1 - pv_1 pv_2}{pv_2(1-pv_1)} \\
&s = \frac{\frac{1}{pv_2} - 1}{\frac{1}{pv_1} - 1}
\end{aligned}
$$

This correction allows us to have the same precision and NPV for both datasets. It is equivalent to considering the precision and NPV in the case the prevalences of both datasets are equal. All stages have be run on the test set, except for the missingness analysis that have been run of the training set.

Appendix C. Main findings from the example report

We will now outline the key takeaways from the fairness audit of the circulatory failure early-warning system (Yèche et al., 2021) that we infer from the sample report (Appendix D). This report was obtained by running FAMEWS on the averaged predictions from 10 LGBM models trained with different random seeds on the HiRID dataset. It can serve as an example of how to interpret such an analysis account.

C.1. Systematic performance discrepancy for male patients

In the summary table **Summarized performance metrics per grouping** (2.1.1.a), we can notice that for almost every metric (except one) the model performs worse on male patients than on female patients. Moreover, in the next aggregated view, it is highlighted that an important part of these metrics is statistically significantly worse. However, looking at the more detailed analysis grouping by sex (section 2.2.1), we realized that the discrepancy in performance (delta value) seems relatively small. The feature ranking doesn't vary significantly between females and males.

C.2. Minority categories aren't always worse off

In the summary tables (from 2.1.1.a to 2.1.1.d) **Summarized performance metrics per grouping**, we can notice that the worst-performing category rarely aligns with the minority category.

C.3. The effect of prevalence correction

If a cohort has a higher prevalence than the others then its performance is decreased by the prevalence correction, while if it has a lower prevalence its performance will be pushed. Thus, it is not surprising to observe that the gap between female and male patients is increased after the correction of AUPRC (Figure 2.2.1.a). In contrast to the effect on neurological patients, where the performance discrepancy in AUPRC has vanished after the correction, as the prevalence of events is the lowest for the neurological cohort (Figure 2.2.3.a). However, one can wonder whether it makes sense to correct for prevalence, i.e. whether we should compare these cohorts under the assumption that they have similar prevalences. It is then important to discuss with clinicians to gain an understanding of how a specific patient attribute impacts the prevalence.

C.4. Label bias for neurological patients

In the **Summary view based on the ratio of significantly worst metrics** (subsection 2.1.1), it is underlined that the worst performance discrepancy over the entire set of cohorts is for neurological patients on event-based recall. They also appear a lot in the table **Top 3 categories with biggest performance metric discrepancies** (2.1.3.a), emphasizing that the model is biased against them.
This is also reflected in the bias of outcomes analysis where neurological patients have, by far, the biggest disparity in the time gap between correct alarm and event (section 3).
The **Medical variable analysis** (section 4) can hint at an explanation for these discrepancies. Indeed, neurological patients have a much higher median value for mean arterial pressure (MAP) than other cohorts (see subsection 4.2.3). This variable is used to construct the label for circulatory failure. We can then wonder whether the label definition is correct for these patients. These results trigger discussions with clinicians in order to adapt the model design and use for neurological patients.

C.5. Dependence of the intensity of measurements on patients' cohorts

We run the **Missingness analysis** (section 6) for arterial lactate (*a_Lac*) and peak inspiratory pressure (*Spitzendruck*). For both of these medical variables, the intensity of measurements is dependent on the patients' groupings, both demographic and clinical. Recall which is a critical metric for our type of application, since we don't want to miss a patient in circulatory failure, is significantly worse when the measurement is missing and the delta values seem quite important. This suggests that missingness has a critical impact on model performance. This sparks processes to improve the imputation strategy and also to dialogue with clinicians in order to gain a better understanding of these patterns of missingness.

Appendix D. Example of the report

A sample report can be found at https://github.com/ratschlab/famews/blob/main/data/sample_reports/hirid_circ_fairness_report.pdf.
In this report, *APACHE group* refers to the admission reason. To understand the meaning of the medical variables, please refer to the data description table of the HiRID benchmark (Yèche et al., 2021): https://github.com/ratschlab/HIRID-ICU-Benchmark/blob/master/preprocessing/resources/varref.tsv.