Information Processing and Management 58 (2021) 102657

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/ipm

FairLens: Auditing black-box clinical decision support systems ☆

Cecilia Panigutti [a,*], Alan Perotti [b], André Panisson [b], Paolo Bajardi [b], Dino Pedreschi [c]

[a] Scuola Normale Superiore, Pisa, Italy
[b] ISI Foundation, Turin, Italy
[c] University of Pisa, Italy

ARTICLE INFO

Keywords:
Clinical decision support systems
Fairness and bias in machine learning systems
eXplainable artificial intelligence

ABSTRACT

The pervasive application of algorithmic decision-making is raising concerns on the risk of unintended bias in AI systems deployed in critical settings such as healthcare. The detection and mitigation of model bias is a very delicate task that should be tackled with care and involving domain experts in the loop. In this paper we introduce FairLens, a methodology for discovering and explaining biases. We show how this tool can audit a fictional commercial black-box model acting as a clinical decision support system (DSS). In this scenario, the healthcare facility experts can use FairLens on their historical data to discover the biases of the model before incorporating it into the clinical decision flow. FairLens first stratifies the available patient data according to demographic attributes such as age, ethnicity, gender and healthcare insurance; it then assesses the model performance on such groups highlighting the most common misclassifications. Finally, FairLens allows the expert to examine one misclassification of interest by explaining which elements of the affected patients' clinical history drive the model error in the problematic group. We validate FairLens' ability to highlight bias in multilabel clinical DSSs introducing a multilabel-appropriate metric of disparity and proving its efficacy against other standard metrics.

1. Introduction

The growing availability of Electronic Health Records (EHR) and the constantly increasing predictive power of Machine Learning (ML) models are boosting both research advances and the creation of business opportunities to deploy clinical Decision Support Systems (DSS) in healthcare facilities (Davenport & Kalakota, 2019; Jiang et al., 2017; Moja et al., 2019). Since many of such models are not equipped to differentiate between correlation and causation, they might leverage spurious correlations and undesired biases to boost their performance. While there is an increasing interest in the AI community to commit to interdisciplinary endeavors to define, investigate and provide guidelines to tackle biases and fairness-related issues (Obermeyer, Powers, Vogeli, & Mullainathan, 2019; Pedreschi, Ruggieri, & Turini, 2008; Saleiro et al., 2018), quantitative and systematic auditing of real-world datasets and ML models is still in its infancy.

☆ DP and CP acknowledge funding from the European Union's Horizon 2020 Excellent Science - European Research Council (ERC) programme under grant n. 834756 "XAI - Science and technology of eXplainable AI decision making" and partial support from the European Union's Horizon 2020 research and innovation programme under grant n. 952026 "HumanE AI Network", grant n. 871042 "SoBigData++: European Integrated Infrastructure for Social Mining and Big Data Analytics" and grant n. 952215 "TAILOR - Foundations of Trustworthy AI - Integrating Reasoning, Learning and Optimization". PB, AP and AP acknowledge partial support from Research Project "Casa Nel Parco" (POR FESR14/20 - CANP - Cod. 320 - 16 - Piattaforma Tecnologica "Salute e Benessere") funded by Regione Piemonte in the context of the Regional Platform on Health and Wellbeing, Italy and from Intesa Sanpaolo Innovation Center, Italy. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

* Corresponding author.
E-mail address: cecilia.panigutti@sns.it (C. Panigutti).

https://doi.org/10.1016/j.ipm.2021.102657
Received 30 November 2020; Received in revised form 18 March 2021; Accepted 26 May 2021
Available online 22 June 2021

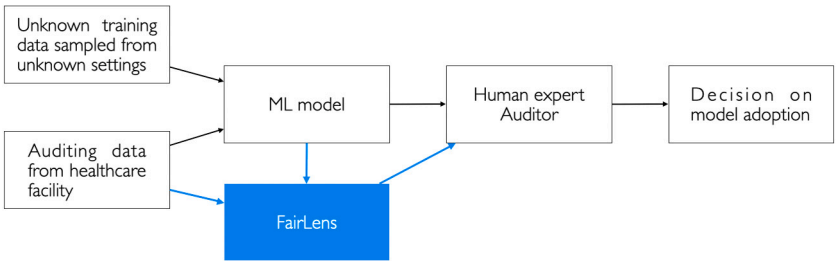

Fig. 1. FairLens as a tool for auditing a clinical decision support system before its deployment in a healthcare facility. Our contribution, represented by the color-filled box, provides to the auditor an instrument to detect and explain systematic ML model biases on protected groups.

Ensuring the fairness of the suggestions provided by ML-based clinical DSS is a delicate task that requires to consider the whole process that goes from data to action. In critical scenarios, ML models do not make autonomous decisions without the supervision of a human; however, they might inadvertently learn to discriminate using unjustified bases for differentiation that reflect a history of systematically adverse outcomes for certain groups (Barocas, Hardt, & Narayanan, 2017; Pedreschi et al., 2008; Pierson, Cutler, Leskovec, Mullainathan, & Obermeyer, 2021), thus leveraging and perpetuating harmful biases in their suggestions. Even under human supervision, the issue of biased suggestions of clinical DSSs is problematic since it has been shown that clinicians are affected by automation-bias, i.e., the tendency to over-rely on automation (Goddard, Roudsari, & Wyatt, 2012; Hillson, Connelly, & Liu, 1995; Lindow, Kron, Thulesius, Ljungström, & Pahlm, 2019). These findings highlight the importance of auditing the clinical DSS before it reaches its end-user.

While the source and the impact of errors of clinical DSS suggestions are numerous, in this paper we focus on errors that lead to systematic biases, and as consequence might cause fairness issues. In other words, we analyze the performance of a ML model across legally recognized protected groups such as gender, ethnicity, age, and on a proxy of socioeconomic status such as healthcare insurance. Indeed, model performance could create fairness issues if the algorithm suggestions on a protected group are systematically wrong (Obermeyer et al., 2019; Seyyed-Kalantari, Liu, McDermott, & Ghassemi, 2020). In this paper, we present FairLens, an auditing tool that allows to test a clinical DSS before its deployment, i.e., before handling it to final decision-makers such as physicians and nurses. The designated user of FairLens is a healthcare facility expert who wants to audit the ML model before adopting and deploying it in the facility, as illustrated in Fig. 1.

In this representation, the origin of model bias might be either in the unknown training data or in the learning process. Bias in training data is usually attributable to a lack of cohort diversity that reflects the data collection process. However, it can also reflect some discriminatory practices (Boag, Suresh, Celi, Szolovits, & Ghassemi, 2018), the historical exclusion of women and ethnic minorities from clinical trials (Heiat, Gross, & Krumholz, 2002; Mason, Hussain-Gambles, Leese, Atkin, & Brown, 2003) or lack of access to care for patients with lower socioeconomic status (Ellis & Jacobs, 2021; McMaughan, Oloruntoba, & Smith, 2020). Since it is generally not possible to access the training data used to build the clinical DSS, FairLens can become a powerful tool to assess if the model is appropriate for the specific hospital's reference population, i.e., the auditing data of Fig. 1. Indeed, FairLens allows the human expert to perform a thorough analysis of potential fairness issues. However, the final decision on whether the signaled bias constitutes a real problem or it is a justified basis for differentiation is left to the auditor. Ideally, the FairLens user is an IT expert with a quantitative background and an in-depth knowledge of the healthcare setting, for example, the director of the IT department of a big hospital. This type of user usually has the responsibility to ensure the quality and trustworthiness of new technologies before adoption. FairLens then becomes an additional tool to understand whether to adopt the system or to evaluate if a bias mitigation strategy is needed, for example, by post-processing the DSS outcomes.

Our Research Question is therefore the following:

> ***How can we audit a black-box Clinical Decision Support Systems in order to detect potential biases on different groups and explain its mislabelings on specific data points?***

FairLens takes bias analysis a step further by explaining the reasons behind the poor model performance on specific groups. FairLens embeds explainability techniques in order to explain the reasons behind model mistakes instead of simply reporting model scores. FairLens first stratifies patients according to attributes of interest such as age, gender, ethnicity, and insurance type; it then applies an appropriate metric to identify patient groups where the model performs poorly. Lastly, it identifies the clinical conditions that are most frequently misclassified for the selected group and explains which elements in the patients' clinical histories are influencing the misclassification.

Throughout this paper, we present a use case where FairLens is used to investigate the potential biases in ML models trained on patients' clinical history represented as diagnostic codes using the *International Classification of Diseases* (ICD) standard. This type of structured data allows for a machine-readable representation of the patient's clinical history and is commonly used in longitudinal ML modeling for phenotyping, multi-morbidity diagnosis classification and sequential clinical events prediction (Che, Purushotham, Cho, Sontag, & Liu, 2018; Choi, Bahadori, Schuetz, Stewart, & Sun, 2016; Xiao, Choi, & Sun, 2018). The implicit assumption behind the use of ICD codes in this kind of ML applications is that these codes are a good proxy for the patient's actual health status.

However, ICD codes can misrepresent such status due to many potential sources of error in translating the patient's actual disease into the respective codes (Chen, Szolovits, & Ghassemi, 2019; O'malley et al., 2005). This is particularly true when ICD codes are fed into *black-box* ML models, i.e., models whose internal decision-making process is opaque.

The presented methodology is designed to be applied to any sequential ML model trained on ICD codes, and we describe the whole auditing process applied to a use-case in this context. In this use-case, we use the most recent update of one of the largest freely available ICU datasets, the MIMIC-IV dataset (Johnson et al., 2020). In this scenario, we show how a domain expert can use FairLens to audit a multilabel clinical DSS (Choi et al., 2016) acting as a fictional commercial black-box model. A ML model is trained with a subset of MIMIC-IV and deployed as a black-box clinical DSS, while the remaining MIMIC-IV data acts as the healthcare facility's historical medical database for auditing. Moreover, we show that FairLens is not limited to this specific setting: a use-case where it is applied to a different clinical decision support system is shown in the Supplementary Information. Finally, we include a computational experiment to validate FairLens in a controlled setting where a known bias is artificially injected in the black-box. In this experiment, a ML model is trained with a biased subset of MIMIC-IV where some categories are under-represented. The model is deployed as a black-box clinical DSS, while a fixed and unbiased MIMIC-IV subset acts as the database for auditing. We compare the results of FairLens using different disparity measures and we show that, as long as the right measure is selected, FairLens effectively detects the synthetically injected bias.

We believe that applied research and quantitative tools to perform systematic audits specific to healthcare data are very much needed in order to establish and reinforce trust in the application of AI-based systems in such a high-stakes domain. FairLens is a first step to make fairness and bias auditing a standard procedure for clinical DSS. We envision such a procedure to be used to monitor bias and fairness issues in all clinical DSSs life-cycle stages: both during model development and training, in the testing phase in controlled real-settings and to constantly monitor the performances over time.

The remainder of the paper is structured as follows: Section 2 provides an overview of ML applications on healthcare data, the fairness problem, and eXplainable Artificial Intelligence. Section 3 introduces FairLens, our novel framework for discovering and explaining group-related disparities, and Section 4 provides a step-by-step application example of FairLens. Section 5 presents a computational experiment to validate FairLens in a controlled setting where a known bias is artificially injected in the black-box. Section 6 ends the paper with a final discussion and directions for future work. The code to run our experiments as well as our results are available on GitHub.[1]

2. Background and related work

Advances in artificial intelligence (AI) in healthcare offer groundbreaking opportunities to enhance patient outcomes, reduce costs, and impact population health (Topol, 2019; Yu, Beam, & Kohane, 2018). Unprecedented results have been achieved leveraging deep neural networks for pattern recognition to help interpret medical scans (Chilamkurthy et al., 2018; Lindsey et al., 2018; Nam et al., 2019; Titano et al., 2018), pathology slides (Bejnordi et al., 2017; Capper et al., 2018; Coudray et al., 2018), skin lesions (Esteva et al., 2017; Haenssle et al., 2018), retinal images (Abràmoff, Lavin, Birch, Shah, & Folk, 2018; Gulshan et al., 2016) and electrocardiograms (Madani, Arnaout, Mofrad, & Arnaout, 2018; Zhang et al., 2018) to name few examples.

The ability to predict key outcomes can also be exploited to improve clinical practice by training DSSs with Electronic Health Records (EHR) (Avati et al., 2018; Chen, Hao, Hwang, Wang, & Wang, 2017; Norgeot, Glicksberg, & Butte, 2019; Rajkomar et al., 2018; Shameer et al., 2017). Compared to more traditional research-oriented clinical data, EHR are usually collected during the clinical encounter and therefore are a more direct reflection of the health status of the population (Casey, Schwartz, Stewart, & Adler, 2016). Among the many types of EHR, the ICD codes are the easiest to process and to be fed into a ML model (Miranda-Escalada, Gonzalez-Agirre, Armengol-Estapé, & Krallinger, 2020; Polignano, Suriano, Lops, de Gemmis, & Semeraro, 2020). ICD stands for *International Classification of Diseases*, which is the standard for the reporting and coding of diseases and health conditions (WHO et al., 2018). ICD codes primary use is to share health information in a structured way. In particular, they are used to share patients' clinical history across hospitals, monitor prevalence of diseases, evaluate hospital performances, and fill the claims for health insurance reimbursement. The ICD codes are assigned to each patient by trained human experts using the information in healthcare providers' clinical notes. The task of translating clinical notes into ICD codes can be made difficult by the use of synonyms and abbreviations that can lead to many ambiguities. Therefore, the process is prone to errors and fraudulent behaviors such as assigning more advantageous codes for reimbursement reasons (O'malley et al., 2005; Piper, 2013) which might create a bias related to patient's socioeconomic status and insurance (Chen et al., 2019). These and many other sources of biases might be present in EHR and can be difficult to track especially if they are fed into a *black-box* ML model.

In general, besides the technological challenges, the various stakeholders involved with the healthcare ecosystem (clinicians, patients/patient advocate, researchers, federal agencies and industry) identified the following urgent priorities for healthcare applications: trustworthiness, explainability, usability, transparency and fairness (Cutillo et al., 2020). As suggested in Raji et al. (2020), before launching (or deploying) a new ML-based product, a thoughtful auditing process is needed. While the auditing process involves multiple stakeholders and embrace several aspects of product development, one of the ultimate goals is to help understanding if the ML model outcomes are fair. Consequently, the auditing process helps to choose the best actions to perform or the best bias mitigation strategy to adopt. Building an auditing system first requires defining fairness according to societal values and then operationalize it. Many efforts have been devoted to detecting and measuring discrimination in model decisions (Hajian,

[1] https://github.com/CeciPani/FairLens.

Bonchi, & Castillo, 2016; Ruggieri, Pedreschi, & Turini, 2010; Zemel, Wu, Swersky, Pitassi, & Dwork, 2013). Several definitions and methodologies have been proposed to measure bias and fairness (Dwork, Hardt, Pitassi, Reingold, & Zemel, 2012; Hardt, Price, & Srebro, 2016; Luong, Ruggieri, & Turini, 2011; Pedreschi et al., 2008); however, despite the effort, a general consensus on such measures is still missing. This is because the most appropriate fairness metric is highly context-dependent. Generally speaking, the most prevalent approach to fairness in machine learning is to solicit for approximate parity of some statistics of the predictions (such as false negative rate) across pre-defined groups (Chouldechova, 2017; Kearns, Neel, Roth, & Wu, 2018; Kleinberg, Mullainathan, & Raghavan, 2016). Moreover, there are very few available general-purposes resources to operationalize them (Adebayo et al., 2016; Bellamy et al., 2018; Saleiro et al., 2018; Tramer et al., 2017). The majority of such research has focused on binary or multi-class classification problems to prevent discrimination based on sensitive attributes assessing fairness issues between only two groups (e.g. female vs. male, black vs. white) (Feldman, Friedler, Moeller, Scheidegger, & Venkatasubramanian, 2015), and a few studies focus specifically on multi-label classification problems, which is the learning problem of the presented FairLens use-case, with many concentrating on fairness in ranking and recommendation systems (Abdollahpouri, Burke, & Mobasher, 2017; Edizel, Bonchi, Hajian, Panisson, & Tassa, 2020; García-Soriano & Bonchi, 2020). In the context of medical applications, a recent paper (Chen et al., 2020) suggested that the post-deployment inspection of model performance on groups and outcomes should be one out of five ethical pillars for equitable ML in the advancement of health care.

Another staple of this paper is the research field of eXplainable Artificial Intelligence (XAI) (Gunning, 2017). Indeed, FairLens embeds explainability techniques to output the fairness report. XAI techniques have the goal to *explain* (i.e., present in human-understandable terms) the decision-making process of an AI system. The need for this kind of technique stems from the fact that the internal decision-making process of many state-of-the-art AI systems is opaque. This can happen either because the source code of the algorithm is proprietary software and cannot be directly inspected, or because the model implements a subsymbolic (numerical) representation of knowledge, often paired with highly non-linear correlations, or both. There are two main approaches to model explanation in the literature (Guidotti et al., 2018): the *transparent-by-design approach* (Angelino, Larus-Stone, Alabi, Seltzer, & Rudin, 2017; Caruana et al., 2015; Wang et al., 2017) and the *post-hoc* approach (Lundberg & Lee, 2017; Panigutti, Perotti, & Pedreschi, 2020; Ribeiro, Singh, & Guestrin, 2016). Methods falling into the *transparent-by-design* category aim to train models that are both interpretable and accurate. Two illustrative examples of this kind of model are Generalized Additive Models (GAM) and Generalized Additive Models with Pairwise Interactions (GAM2) (Caruana et al., 2015). Indeed, once trained, the user can directly inspect the knowledge learned by these models visualizing the relationship between the output and a single feature (in the case of GAM) or between the output and a pair of features (in the case of GAM2). Another representative example of a transparent-by-design model is Bayesian Rule Lists (Letham, Rudin, McCormick, Madigan, et al., 2015), where the trained model consists of an ordered list of if-then rules that describe the decision-making process of the model. Generally, these transparent-by-design models are based on models recognized as inherently interpretable in the literature: linear models, decision trees, and if-then rules. While this approach to model explanation is always ideal, it is not applicable in all scenarios. Building transparent models with competitive prediction performance is particularly difficult in the case of multi-class and multi-label classification problems (Zhang et al., 2019). Furthermore, this approach to model explanation cannot be applied when the final goal is to audit the decision-making process of a proprietary software, which is the case presented in this paper.

In this perspective, a greater flexibility is offered by the *post-hoc* approach. Indeed, this approach focuses on extracting explanations from a black-box model after training. Several methods falling into the post-hoc category are *model-agnostic*, i.e., they can be applied to any black-box since they analyze only its input–output behavior (Lundberg & Lee, 2017; Ribeiro et al., 2016). On the other hand, *model-aware* post-hoc XAI techniques are often based on specific ML models and require access to internal values of the black-box such as the gradients in the convolutional layers for GradCam (Selvaraju et al., 2017) or the *attention scores* (Vaswani et al., 2017) as discussed in Wiegreffe and Pinter (2019). Since the *model-agnostic* approach to model explanation focuses only on its input–output behavior, a plethora of methods have been developed to deal with a variety of data sources (relational Anjomshoae, Kampik, & Främling, 2020; Guidotti et al., 2018; Panigutti, Guidotti, Monreale, & Pedreschi, 2019; Ribeiro, Singh, & Guestrin, 2018, text Mullenbach, Wiegreffe, Duke, Sun, & Eisenstein, 2018, images Guidotti, Monreale, Matwin, & Pedreschi, 2020; Selvaraju et al., 2017, sequences Panigutti et al., 2020 or several of them Lundberg & Lee, 2017; Ribeiro et al., 2016), and learning problems (binary and multi-label classification, regression, scoring) allowing the user to choose the best explainer for the task at hand. These models are also often local, which means that the provided explanations are valid only for individual predictions and fail to generalize to the whole model's logic. To overcome this limitation, some new XAI methods have been proposed to generalize the local explanations combining them into a surrogate model able to mimic the black-box logic while being interpretable at the same time (Lundberg et al., 2020; Setzu, Guidotti, Monreale, & Turini, 2019; Setzu et al., 2021). FairLens methodology, as explained in more details in Section 3, can be applied using any explainer that allows to combine single explanations into a global one.

3. Fairlens pipeline

This Section describes the FairLens methodology to audit black-box Clinical Decision Support Systems in order to (*i*) detect potential biases on different groups and (*ii*) explain its mislabelings on specific data points.

Here we describe an end-to-end use of FairLens on a specific setting (i.e. prediction of future health conditions, based on past observation of ICD codes), and we provide an alternative scenario in the Supplementary Information. Indeed, it is worth stressing that the functional blocks of the pipeline are quite general and thus FairLens can also be used in different settings after an appropriate tailoring of the modules. In particular, different applications might be interested in stratifying the data points according to different categories other than gender, ethnicity, age and insurance. Moreover, according to the classification problem at hand, the

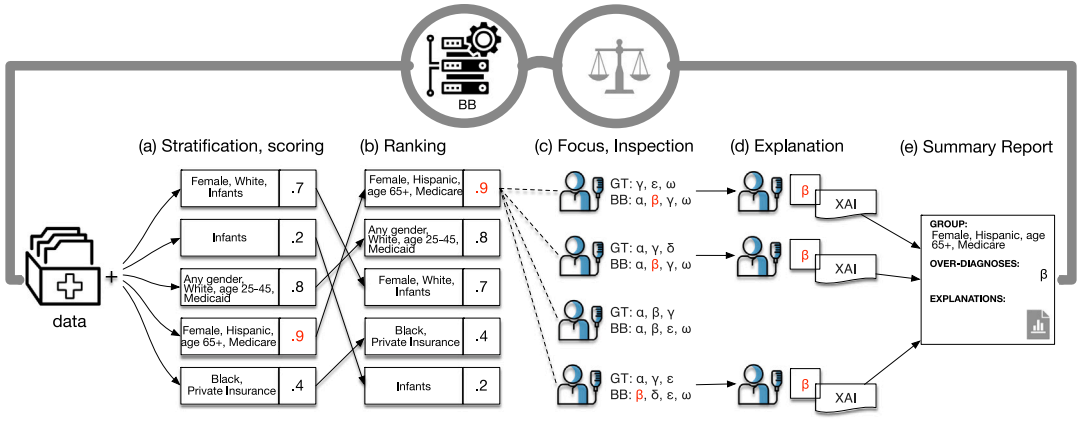

Fig. 2. FairLens pipeline: a tool to support human experts investigating if a black-box clinical decision support system behaves differently on groups based on protected attributes, highlighting which health conditions are more often misclassified and why.

scoring measure might be different from the one presented here for the high-dimensional multi-label classification, and clearly the explanation method should be suitable for the black-box as well. Such considerations highlight the potential of FairLens as a useful framework to allow humans inspecting algorithmic decision-making pipelines, without delegating to yet another automated tool the delicate task of auditing unintended and potentially harmful consequences of decision support systems. As such, our approach provides insights about the *who* and the *why* of the differential treatment of a clinical decision support system on certain groups, letting the human experts understanding if such behavior is legit or may lead to fairness issues.

Given a black-box to audit, the building boxes of the pipeline described hereafter are: stratification, scoring, ranking, inspection, explanation and summary report. A bird's-eye view of the pipeline is depicted in Fig. 2.

Let BB be a sequential black-box ML model trained on ICD data. The model can be available as an on-premise-installed software or it could be integrated via an exposed API. The only requirement about BB is that it can be queried at will.

Let $P = \{p_1, \dots, p_N\}$ be the set of patients. Let each patient p_i be represented as (p_i^{att}, p_i^{ch}), where p_i^{att} is a set of attributes such as *ethnicity*, *gender*, and *insurance type*, and $p_i^{ch} = \{v_{i,1}, \dots, v_{i,V}\}$ is the clinical history represented as a sequence of visits. In turn, each visit is represented by a set of ICD codes.

Let $v_{i,j}^{BB} = BB(\{v_{i,1}, \dots, v_{i,j-1}\})$ be the prediction of the black-box for the jth visit of patient p_i.

It is worth to notice that the p_i^{att} are not part of the feature space of the BB, and in principle the patients' attributes could be more than those presented here to exemplify the use of FairLens. In general, p_i^{att} could include any attribute that is not used by the model to predict future health conditions, but can be collected in a structured database (e.g. education level, job status).

3.1. Stratification

The first step of our methodology is depicted in Fig. 2(a). Since we aim to compare the ML model performance across groups, we stratify our patients set P according to a set of conditions c on the set of attributes p^{att}, e.g. $c = \{\text{age} \leq 40, \text{insurance} = \text{Medicaid}\}$. Given a set c of conditions, we define a *group* G as the set of non-first visits of each patients whose attributes match the conditions in c:

$$G_k = \{v_{i,j} \mid j > 1, v_{i,j} \in p_i^{ch}, p_i \in P, p_i^{att} \in c_k\}$$

The stratification process produces a set of groups $G_1, \dots, G_M$. While the stratification process is based on the attributes of patients, we create different data-points for each non-first visit, so that we can evaluate the performance of the model on every visit of the patients' clinical history. Considering each visit as a different data point is necessary because some demographic attributes might change between two visits of the same patient (consider for example *age* and *healthcare insurance*). The first visit of each patient ($j = 1$) are excluded because in those cases the model has no previous patient history to base its prediction upon.

We remark that there is a degree of freedom regarding which set of attributes are considered. The granularity might be tuned at will, ranging from one-attribute constraints $\{\text{gender} = \text{F}\}$ to more detailed constraints $\{\text{gender} = \text{F}, \text{age} \geq 65, \text{ethnicity} = \text{white}, \text{insurance} = \text{Medicare}\}$.

A domain expert might suggest specific condition sets to isolate a given sub-cohort of known interest, whereas a technician might opt for building a lattice of all possible combinations of constraints. The attributes considered here are deemed relevant for auditing purposes as existing literature suggests that minority groups might be at risk of fairness issues, and protected attributes (i.e. traits or characteristics that, by law, cannot be discriminated against as age and gender) should not affect the model performance. Here, we also considered the insurance type as it is a proxy for socioeconomic status. According to data availability, other attributes could be further added to the stratification process. We also remark that some patients might not occur in any group or occur in more than one, depending on the provided conditions.

3.2. Scoring

After the stratification step, FairLens proceeds to the scoring phase. For every non-first visit $v_{i,j}$ occurring in any group G_k, we query the BB on the previous clinical history of that patient, so that we can compare the ground-truth visit $v_{i,j}$ with its predicted counterpart $v_{i,j}^{BB} = BB(\{v_{i,1}, \dots, v_{i,j-1}\})$. We therefore obtain the predicted counterparts for every visit in every group, and we can evaluate how different groups fare in terms of truth-prediction disparity.

Although many works in literature define disparity as a distance according to a reference group (Keppel et al., 2005), here we choose to define disparity as a measure relative to a target standard, that in the case of ML algorithms might be e.g. perfect prediction of the target values. Therefore, for the purposes of this discussion, we propose the following definition of disparity:

> *The quantity that separates a group from a target standard using a particular measure of performance.*

Hence, a *disparity function* $d : G_k \to s_k$ maps every group G_k to a *disparity score* s_k.

FairLens includes a number of disparity functions, such as the standard classification metrics (such as accuracy and F1-score) and distribution-comparison functions like the Wasserstein distance. Custom disparity functions can be used, as long as their results can be used for ranking. Given a disparity function, FairLens computes the score s_k for each group G_k, which represents the performance of the BB on that specific set of patients.

3.3. Ranking

Once each group has been scored, FairLens ranks the groups, as depicted in Fig. 2(b). The ranking highlights groups where the BB performs relatively poorly, signaling them to domain experts for further inspection. Alternatively, the domain experts might arbitrarily select one group for further inspection, regardless of their scores, due to the cohort's known peculiarities or clinical-dependent reasons.

3.4. Inspection

Given a specific group G_k flagged for further inspection by the group ranking function, FairLens compares the black-box prediction $v_{i,j}^{BB}$ with the ground truth $v_{i,j}$ for each visit in G_k. The goal of this step is to check for systematic bias of the BB on the group of patients. For each diagnostic code, the relative frequencies in the predicted and true values are computed and we define the *misdiagnosis score* the difference between these two values. Ranking the codes by misdiagnosis scores allows to highlight which diagnostic codes are particularly over- or under-predicted (high and low difference values respectively). FairLens thus displays the top three over- and under-represented codes to the domain expert who can ask for an explanation for the highlighted conditions that might result in producing or reinforcing systematic over- or under-treatment. In Fig. 2(c), we have labeled the true visit value as GT (for *ground truth*); in the mock example it can be observed that the code β is over-represented.

3.5. Explanation

In order to extract an explanation for the mislabeled code, FairLens first assigns binary labels on the visits of the group of interest. Suppose the domain expert wants to understand what elements of the group clinical histories are most influencing the over-representation of ICD code β in the inspected group G_k, then at each visit $v_{i,j} \in G_k$ will be assigned a binary label representing the misclassification of the ICD code β:

$$l(v_{i,j}) = \begin{cases} 1 & \text{if } (\beta \in v_{i,j}^{BB}) \oplus (\beta \in v_{i,j}) \\ 0 & \text{if } (\beta \in v_{i,j}^{BB}) == (\beta \in v_{i,j}) \end{cases}$$

Then, FairLens selects all the misclassified visits (binary label 1) and explains them using a local XAI technique for sequential healthcare data. Typically XAI techniques are used to explain the outcome of a black-box ML model. In this setting, we want to explain why the specific code was wrongly assigned, and we do so by providing the XAI technique with the custom binary label.

More generally, we define the *Explainer* as a function:

$$\xi : (BB, x_i, \beta) \to \{ f_1 \gtrless t_1 \dots, f_F \gtrless t_F \}$$

that maps a blackbox BB, a patient's feature vector x_i and a clinical code β to a set of decision rule premises $\{ f_1 \gtrless t_1 \dots, f_F \gtrless t_F \}$ where each f is a feature in x_i that, in combination with a threshold value t, explains why BB misclassified β for the patient p_i. In the case where a black-box BB predicts β from a feature vector x_i that is the patient's clinical history p_i^{ch}, the feature names f are a subset of the medical codes in p_i^{ch}.

It is worth noting that while XAI techniques are usually employed to explain the reasons behind a black-box decision, thanks to the aforementioned binarization process, FairLens uses them to explain the reasons behind a specific mislabeling. Furthermore, we observe that when a model-agnostic XAI technique is employed, FairLens can be used audit any model without having access to its internal structure or parameters. However, FairLens can be used with model-aware XAI techniques too, and we provide an example in the Supplementary Information.

Table 1
MIMIC-IV: Data from patients with at least two hospital admissions.

Number of patients	43,697
Number of admissions	164,411
avg. nr. of admissions per patient	3.76
max nr. of admissions per patient	146
Number of unique ICD-9 codes	8259
avg. nr. of codes per admission	11.22

3.6. Reporting

Finally, FairLens combines the local explanations of each mislabeled visit of group G_k in one set of *global* rules; this corresponds to step (e) in Fig. 2. The local explanations extracted by FairLens are in the form of decision rules with premises. Each condition of the rule premise follows the pattern

$$\text{ICD\_code} \gtrless threshold\_value$$

where the *threshold value* expresses whether and when the ICD code was observed in the patient's clinical history. These local explanations are merged by FairLens employing a state-of-the-art XAI technique, GlocalX (Setzu et al., 2021), that outputs a compact set of global rules by hierarchically merging the local explanations based on their similarity. Finally, FairLens translates the final set of global rules into natural language and presents the report to the user.

4. Use case: auditing a medical decision support system

In this section we show how a domain expert can use FairLens on the historical data available at her healthcare facility to audit a fictional commercial clinical *decision support system* (DSS) that predicts patient's future clinical events based on their clinical history. We assume that the domain expert has access to the DSS as a *black-box*, i.e. she can query the DSS at will but has no access to its source code, to its weights or to the data used for its training. We use the MIMIC-IV (see Section 4.1) database of electronic health records as the fictitious historical database of the facility and DoctorAI (see Section 4.2) as the fictional clinical DSS. We split the dataset in training (29.714 patients, 68%), validation (5.244 patients, 12%) and test set (8.739, 20%). Training and validation sets are used to deploy DoctorAI as a black-box and are not seen during the auditing process, while the patients in the test set are used as auditing data. We exploit DoctorXAI (Section 4.3) as the backbone of the FairLens explainer, and we show how this auditing process is effective to detect and explain potential biases on different groups.

4.1. Dataset: MIMIC-IV

The MIMIC (Medical Information Mart for Intensive Care) (Goldberger et al., 2000; Johnson et al., 2016) database is a single-center freely available database containing de-identified clinical data of patients admitted to the ICU (intensive care unit) of the Beth Israel Deaconess Medical Center in Boston. Its most recent update, MIMIC-IV (Johnson et al., 2020), contains information of 383,220 patients collected between 2008 and 2019 for a total of 524,520 hospital admissions. The database includes patient's demographics, clinical measurements and diagnoses and procedures codes of each admission. We focused our analysis on hospital admissions coded with ICD-9 billing codes and on patients having at least two admissions to the hospital, reducing the number of patients to 43,697 and the number of admission to the hospital to 164,411 (see Table 1).

4.2. Clinical DSS: Doctor AI

Doctor AI by Choi et al. (2016) is a Recurrent Neural Network (RNN) with Gated Recurrent Units (GRU) that predicts the patient's next clinical event's time, diagnoses and medications. For the purpose of this use-case, we focused only on diagnoses prediction. We trained the model on MIMIC-IV using the training and validation set as defined previously using default hyperparameters.

Doctor AI can be trained to predict patient's future clinical event in terms of either CCS (Clinical Classifications Software) or ICD codes. CCS codes are used to group ICD codes into smaller number of clinically meaningful categories. As suggested in Choi et al. (2016) we trained Doctor AI to estimate the probability that a CCS code is assigned to a visit at time $t+1$ given the ICD-9 codes assigned to patient's visits until time t, and measured its performance using Recall@n with $n = 10, 20, 30$.

4.3. Local explainer: DoctorXAI

DoctorXAI (Panigutti et al., 2020) is a post-hoc explainer that can deal with any multi-label sequential model. Since it is agnostic w.r.t. the model, i.e. it does not use any of its internal parameter in the explanation process, it is suitable for our methodology which considers the clinical DSS as a black-box. Furthermore, DoctorXAI exploits medical ontologies in the explanation process and in our case we exploited the ICD-9 ontology. The explanations provided by DoctorXAI are *local* decision rules, which means that they provide the rationale for one particular classification.

In our scenario, we want to provide an explanation for a over- or under-diagnosis observed in a group of patients, therefore FairLens binarizes the black-box probability estimates and it combines the explanations as described in the *Explanation* and *Reporting* paragraphs of Section 3.

Table 2
Clinical DSS performance.

BB recall	@10	@20	@30
On auditing data	0.481	0.623	0.712

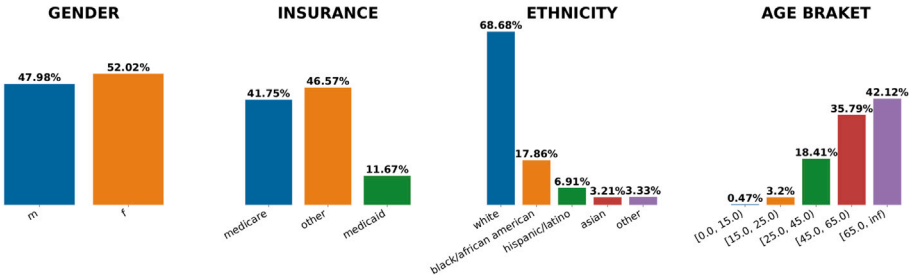

Fig. 3. Distributions of demographic attributes in the auditing data.

4.4. Local-to-global approach: GlocalX

GlocalX (Setzu et al., 2021) is a model-agnostic XAI algorithm that explains the global behavior of black-boxes by aggregating a set of local explanations in the form of decision rules. GlocalX hierarchically merges local explanations optimizing both the complexity and fidelity of the decision rules set, i.e., its size and ability to mimic the black-box behavior correctly. In our case, we used GlocalX to merge all the local explanations extracted by DoctorXAI to explain the individual misclassifications of a group. We stress that while GlocalX is a methodology to generate a transparent model able to mimic the black box's global behavior, in our scenario, we use it as an aggregator of explanations for the patients of the group under investigation, i.e., all the patients having a specific misclassification. Therefore the validity of the provided global explanation is limited to the black-box behavior on those patients. As described in Section 3, DoctorXAI produces rules that follow the pattern ICD_code $\gtrless threshold\_value$, and GlocalX preserves this structure. To map back these rules onto human-readable sentences, we simply revert DoctorXAI's temporal encoding. In order to circumvent the temporal nature of medical history data, DoctorXAI exploits a fairly straightforward temporal encoding, where each ICD9 code receives an exponentially decreasing value according to its occurrence (or lack thereof) in the visits of the patient, explored backwards. The last visit corresponds to a score of .5, the second-to-last to a score of .25, and so on. For instance, if some condition C was diagnosed in the third-to-last and second-to-last visits, but not in the last one, C would be given the value of .375. Given this logic, it is trivial to interpret the inequalities produced by DoctorXAI and aggregated by GlocalX: $C < .5$ means, for instance, that the ICD9 C was not diagnosed in the last visit, while $C \geq .25$ means that the ICD9 C was diagnosed at least once in the last two visits of the patient.

4.5. Auditing DoctorAI on MIMIC-IV

4.5.1. Assessing the dss performance on the healthcare facility data.

The first step that a domain expert would perform before deploying the clinical DSS on her dataset is to measure its global performance on the facility data. In our scenario, a domain expert would obtain the results in Table 2.

4.5.2. Identifying problematic groups of patients.

Once the global performance has been assessed, the domain expert can apply FairLens to discover potential biases learned by the model. The domain expert would start by deciding which attributes to use to stratify the patients. For the purpose of our fictional scenario, we consider the following attributes occurring in the auditing data: *Gender*, *Ethnicity*, *Age* and *Insurance type*. The distributions of these attributes is shown in Fig. 3.

Once these attributes are selected, FairLens computes the disparities across groups. In our scenario, the black-box is a sequential multi-label model that predicts the set of codes diagnosed in the next visit in terms of CCS codes. In this multi-label case, the disparity is evaluated using the *Wasserstein distance* which has already been successfully employed as a loss function for multi-label and multi-class ML tasks (Frogner, Zhang, Mobahi, Araya, & Poggio, 2015) and to post-process the output of a classifier to achieve fair treatment (Jiang, Pacchiano, Stepleton, Jiang, & Chiappa, 2020). This metric measures the distance between two probability distributions: for each group of interest, the distance between the distribution of CCS codes in the black-box output and the same distribution in the ground truth. In our scenario, the DSS outputs the top 30 CCS codes ranked by estimated probability. Similarly to the *recall@k* we define the *disparity score@k* which is the Wasserstein distance between the ground truth and the predicted probability distributions over the top-k CCS codes. From now on, we will perform the analysis using the *disparity@30* unless otherwise specified.

The domain expert can decide to either explore a specific group of interest or to have a comprehensive view of the biases of the DSS on all possible groups.

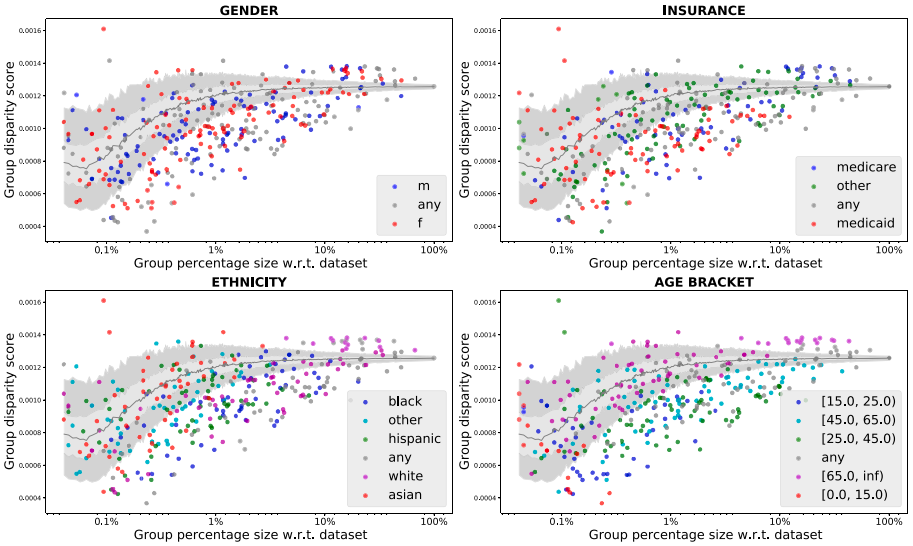

Fig. 4. Normalized disparity scores vs. group sizes with bootstrap outliers bands capturing 50% (light gray) and 95% (dark gray) of the random variability for that group size. The median of the bootstrap distribution is shown as a solid gray line.

Table 3
Groups with the highest disparity score in each group size bin. All disparity scores marked with ∗ are above the 95th percentile of random variability for the group size.

Group size bin	Insurance	Gender	Age group	Ethnicity	Disparity score	Group size
10–50	medicaid	f	25–45	asian	1.00 ∗	23
50–100	medicare	f	over 65	other	0.84 ∗	70
100–200	any	f	over 65	other	0.84 ∗	111
200–400	any	any	over 65	asian	0.88 ∗	286
400–800	any	any	any	asian	0.83 ∗	657
800–1500	other	m	over 65	white	0.86 ∗	1082
1500–3000	medicare	m	over 65	white	0.86 ∗	2783
3000–5000	any	m	over 65	white	0.86 ∗	3894
5000–24 446	medicare	any	over 65	white	0.86 ∗	5679

The scatter plots in Fig. 4 confront the normalized disparity score with the group size for all possible groups. Each scatter plot focus on a specific attribute, and each point represents a group with a combination of attributes, for a total of 340 combinations. The color-coding allows to explore the disparities of each intersectional identity. Data points labeled and color-coded as *any* correspond to groups that do not represent a specific value for the stratification feature: for instance, the group *(male, medicare)* includes patients of all ages and ethnicities.

In the same plots we show the variability in disparity score as function of the group size, when selecting the same number of patients independently of their group assignment. In particular, we randomly sample with replacement 1000 times for each group size to estimate its disparity score's sampling distribution. The plots show the median (solid gray line) and the bands capturing the 50% (light gray) and 95% (dark gray) of the distribution for each group size. The groups falling above the dark gray band's upper limit have a disparity score above the 95th percentile of the distribution for that group size when no demographic variable is considered. These groups are also marked with an asterisk in Table 3.

A higher variability in terms of disparities is observed among smaller groups. While this might suggest fairness issues for relatively rare groups, given the small size of these groups, the high variability and dispersion away from the mean could also occur by chance; therefore Table 3 also provides an overview of the groups with the highest disparity in predefined group-size bins. The results in this table (supported by the Age Bracked plot in Fig. 4) suggest that the DSS seems to often misdiagnose older patients; indeed they are the most prevalent age group with the largest disparity score by group size bin.

4.5.3. Identifying systematic sources of error in the selected group.

For each group, FairLens then computes the misdiagnosis score of each CCS code by subtracting its ground truth value (clinical conditions) from the value predicted by the DSS. This score allows to rank the codes, so that the most over- and under-diagnosed CCS codes can be isolated. Table 4 reports the top 3 groups by disparity score in the largest bins, and the top 3 codes ranked by over- and under-diagnosis scores.

The domain expert auditing the system can further select a specific group for a more in-depth investigation. Suppose she decides to focus on one of the groups with the highest disparity and also a fairly high group-size, for example patients of Asian ethnicity

Table 4
Groups ranked by disparity scores and most over/under-diagnosed conditions when auditing the black-box.

Group	Size	Disp. score	Over-diagnosed	(Misdiagnosis score)	Under-diagnosed	(Misdiagnosis score)
Female, 65+, Medicare, Other ethn.	70	0.83	106: Dysrhythmia 98: Essential hypertension 259: Unclassified	0.027 0.02 0.019	2621: E Codes:Place of occurrence 2603: E Codes: Fall 210: Systemic lupus erythematosus	−0.010 −0.009 −0.007
Female, 65+, Other ethn.	111	0.84	259: Unclassified 98: Essential hypertension 106: Dysrhythmia	0.024 0.024 0.022	2621: E Codes:Place of occurrence 2603: E Codes: Fall 250: Nausea/vomit	−0.009 −0.007 −0.006
Asian, 65+	286	0.88	259: Unclassified 98: Essential hypertension 663: Hist. of mental healthand subs. abuse	0.023 0.020 0.016	6: Hepatitis 204: Other non-traumaticjoint disorder 96: Heart valve disorders	−0.009 −0.008 −0.007

Table 5
Set of rules produced by DoctorXAI and aggregated by GlocalX to explain why the CCS code 96: *Heart valve disorders* was under diagnosed for over-65 Asian patients by the model DoctorAI. Each row group is a rule with a set of premises, each premise is in the form of ICD-9 $\gtrless$ *threshold_value*. For the human-readable description of each ICD-9 code the reader can consult http://www.icd9data.com/.

427.31 ≤ 0.25	**410.91** ≤ 0.25	**396.3** ≤ 0.25	**410.71** ≤ 0.25	**424.0** > 0.25	**162.3** > 0.125
424.1 ≤ 0.25 **V10.3** > 0.004	**425.4** > 0.16 **V49.86** > 0.024	**202.10** > 0.0005 **V12.72** > 0.033	**427.31** > 0.5 **E930.7** > 0.016	**244.9** > 0.5 **V45.82** > 0.244	**E933.1** > 0.1
427.31 > 0.62 **362.50** > 0.125 **V04.81** > 0.25	**V45.82** > 0.125 **530.81** > 0.375	**428.0** > 0.437 **411.1** > 0.25	**567.29** > 0.125 **412** > 0.187	**575.4** > 0.125 **401.9** > 0.25	**574.00** > 0.062 **564.00** > 0.062
427.31 ≤ 0.25					
424.1 > 0.25 **288.3** > 0.0625	**V12.71** > 0.344 **255.9** > 0.0625	**401.9** > 0.148 **V13.01** > 0.25	**305.1** > 0.219	**E849.9** > 0.023	**403.90** > 0.344

and over 65 years of age (see Table 4). This analysis tells the domain expert that across groups the DSS tends to over-diagnose general conditions such as *Essential hypertension* or *Unclassified*. More interestingly, for the group of patients of Asian ethnicity and over 65 years of age, the DSS seems to under-diagnose *Heart valve disorder*, which is a potentially severe condition that might need surgery.

4.5.4. Obtaining explanations for systematic misclassifications.

Once the groups with the highest disparities are identified, the domain expert can use FairLens to obtain an explanation for one particular misclassification. Consider, for example, the under-diagnosis of *Heart valve disorders* (CCS code 96) for over-65 Asian patients. FairLens uses DoctorXAI to discover which elements in the patients' clinical history drive the under-diagnosis of that specific CCS. This is done by first projecting the black box's multi-label output on the single label 96 (as explained in Section 3), then calling DoctorXAI to explain the binarized outcome for the 19 patients where the CCS code 96 was wrongly not diagnosed.

By doing so we obtain 19 explanations, one for each CCS-96-misdiagnosed patient in our patients group. As a further step, the GlocalX local-to-global algorithm aggregates these local explanations into a more compact and doctor-readable global explanation, as introduced in Section 2. GlocalX, for this explanation set, produces the global rules of Table 5. While the original rule set had 19 rules of mean length 10, the resulting rule set contains only 5 rules of mean length 8. Clearly, this is a more compact set but not yet comprehensible.

As a very first feedback to the expert, FairLens produces Fig. 5: this plot highlights the ICD9 codes that occur in the global rules (and therefore are brought out by the FairLens pipeline as misclassification culprits) and are also most common among the patients of the group under scrutiny. In our case, for instance, the domain expert can immediately observe that the highlighted ICD9 codes are 410.91 (*Acute myocardial infarction of unspecified site, initial episode of care*), 396.3 (*Mitral valve insufficiency and aortic valve insufficiency*) and 410.71 (*Subendocardial infarction, initial episode of care*). Fig. 5 provides useful preliminary insights to the FairLens user, but at the same time the information conveyed by the global explanations is richer and can be presented in greater details. First, we want to translate these rules back into natural language, and we do so as explained in the previous subsection: for instance, the last global conjunct is **V13.01**>0.25 and it corresponds to *'Personal history of urinary calculi' was diagnosed in the last visit*. Second, we want to rank our global rules. To do so, we measure the coverage of each rule as the number of patients whose features do not violate the rule, and we select the rules in a greedy fashion, highlighting those with higher coverage. For our case-study, the re-interpreted output of GlocalX is the following:

- FairLens focused on 19 patients
- 13 patients were misdiagnosed because *'Atrial fibrillation'* was not diagnosed in the last visit.
- 5 remaining patients were misdiagnosed because *'Aortic valve disorders'* was not diagnosed in the last visit, *'Other primary cardiomyopathies'* was diagnosed at least once in the latest two visits, *'Mycosis fungoides, unspecified site, extranodal and solid*

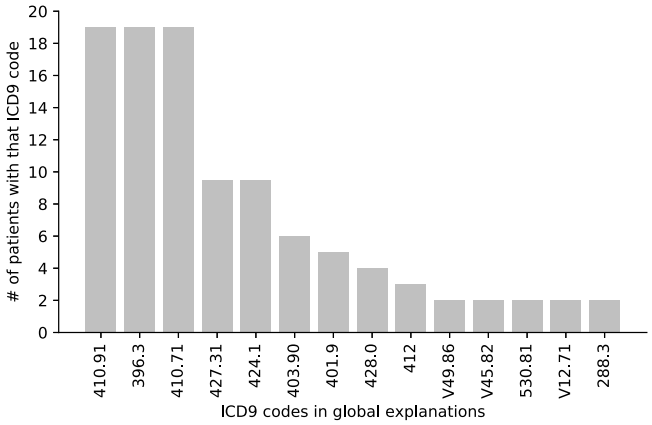

Fig. 5. Aggregated visualization of the relevant ICD-9 codes for the under-diagnosis of *Heart valve disorders* in over-65 Asian patients.

organ sites' was diagnosed at least once in the latest three visits, *'Atrial fibrillation'* was diagnosed in the last visit, *'Unspecified acquired hypothyroidism'* was diagnosed in the last visit, *'Antineoplastic and immunosuppressive drugs causing adverse effects in therapeutic use'* was diagnosed at least once in the latest three visits, *'Personal history of malignant neoplasm of breast'* was diagnosed at least once in the latest three visits, *'Do not resuscitate status'* was diagnosed at least once in the latest three visits, *'Personal history of colonic polyps'* was diagnosed at least once in the latest three visits, *'Antineoplastic antibiotics causing adverse effects in therapeutic use'* was diagnosed at least once in the latest three visits, and *'Percutaneous transluminal coronary angioplasty status'* was diagnosed at least once in the latest two visits.

- 1 remaining patient was misdiagnosed because *'Aortic valve disorders'* was diagnosed in the last visit, *'Hypertensive chronic kidney disease, unspecified, with chronic kidney disease stage I through stage IV, or unspecified'* was diagnosed in the last visit, and *'Eosinophilia'* was diagnosed at least once in the latest three visits.

This human-readable snippet is the final output of FairLens pipeline – it provides medical experts with insights on why the medical decision support system misdiagnosed patients of the selected group, failing to diagnose the highlighted condition, CCS 96 – *Heart valve disorder*.

5. Validation

To empirically validate the reliability of FairLens in discovering biases, we created an artificially biased DSS and we ran the FairLens pipeline on it. The aim of this validation is to check whether the disparity measure used by FairLens is able to highlight the bias we injected in the DSS even when standard measures of multi-label performance (e.g. *recall@n* and *microAUC)* do not detect it.

5.1. Creating the biased dss.

One of the most common causes of bias in machine learning is the under-representation of some categories in the training set. We then performed a random undersampling of patients having *Other* as Insurance, removing 90% of them from MIMIC-IV dataset (sampling A of Fig. 6). Finally, we used this skewed dataset as the training set for DoctorAI creating the biased DSS. While in this case we used such approach to validate the proposed pipeline, it is worth to notice that several studies suggest that ICD9 codes might be severely biased by the insurance type variable (Geruso & Layton, 2020; Harrington, Allen, & Ruchala, 2007; Lyon et al., 2011; Piper, 2013).

The biased DSS created using this training set also contains, by construction, all the biases already present in the original dataset. To check whether the bias detected by FairLens in the biased DSS is actually the one we synthetically injected rather than the one already present in the original dataset, we created a *baseline* DSS by training DoctorAI on a random undersampling of MIMIC-IV (sampling B in Fig. 6). This sampling creates a training set that has the same size as the biased one, but that has the same distributions of demographic variables as the auditing dataset. Fig. 6 shows the resulting distributions of training set demographic variables for the two sampling and for the test set.

The fact that the size of the training set is the same for both the biased and the baseline DSS allows a fair comparison of the performance metrics among the two. Indeed comparing the performance of the biased DSS with a baseline trained on a MIMIC-IV dataset without sampling would result in a baseline performance higher than the biased one only due to the bigger size of training set, creating a confounding factor for the analysis. The performance of the two DSSs on the test set are shown in Table 6.

Comparing the distributions of demographic variables of these two black-boxes (Fig. 6), we note that by removing 90% of patients having *Other* insurance, we also changed the distributions of other demographic variables. Consider, for example, the age distribution in the biased training set. We can see that patients having age 0–15 almost disappear from the dataset.

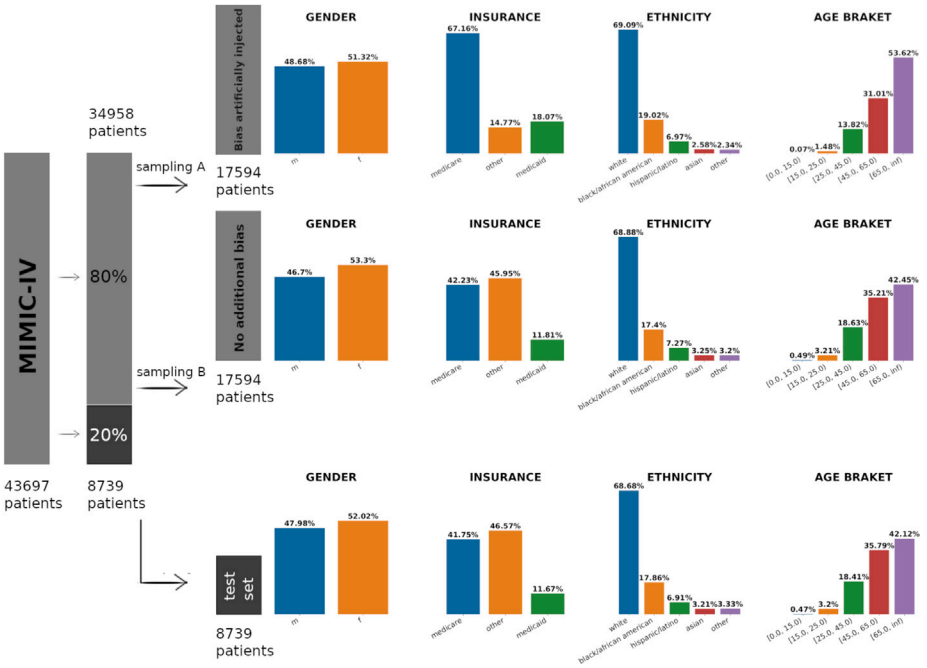

Fig. 6. Sampling procedure and distribution of demographic variables in the training sets and test set. We first extract from the whole MIMIC-IV a test set with 20% of data points. From the remaining points, we extract two training sets with different sampling procedures (sampling A and sampling B). Sampling A produces a training set with artificially injected bias. Sampling B produces a training set with random sampling that respects the same distribution of demographic variables as the original dataset.

Table 6
Performance of clinical DSS trained on the biased and on the baseline training sets.

BB recall trained	@10	@20	@30
On biased training set	0.449	0.586	0.671
On baseline training set	0.454	0.591	0.683

Table 7
Groups with the highest disparity score in each group size bin for the biased DSS. All disparity scores marked with ∗ are above the 95th percentile of random variability for the group size.

Group size bin	Insurance	Gender	Age braket	Ethnicity	Disparity score	Group size
10–50	other	f	0.0–15.0	white	1.00 ∗	10
50–100	other	any	0.0–15.0	any	0.66 ∗	57
100–200	other	m	any	asian	0.42 ∗	149
200–400	any	any	over 65	asian	0.40 ∗	286
400–800	any	any	any	asian	0.39	657
800–1500	other	m	any	black/african american	0.40 ∗	866
1500–3000	medicare	f	over 65	white	0.41 ∗	2896
3000–5000	medicare	f	over 65	any	0.41 ∗	3832
5000–24 447	any	f	over 65	any	0.40 ∗	5351

5.2. Fairlens analysis

We then proceed to run FairLens Pipeline on these two DSS. The first step is to identify potentially problematic groups of patients using FairLens scatterplots (Fig. 7) and tables (Table 7).

Comparing the two scatterplots we can immediately see that FairLens detect both the *Insurance* and the *Age* bias synthetically injected in the biased DSS. Indeed, the majority of patients having the biggest disparity scores are those of age 0–15 and those having insurance *Other*. This is visible also in the tables that show the highest disparity scores binned by group size (see Tables 3 and 7).

We also compared FairLens average ranking aggregated by insurance type for both the biased and the baseline DSS. The results reported in Table 8 show that, for the baseline DSS, FairLens ranks *Medicare* as the insurance having the highest disparity score across different groups, while *Other* is ranked above the others for the biased DSS.

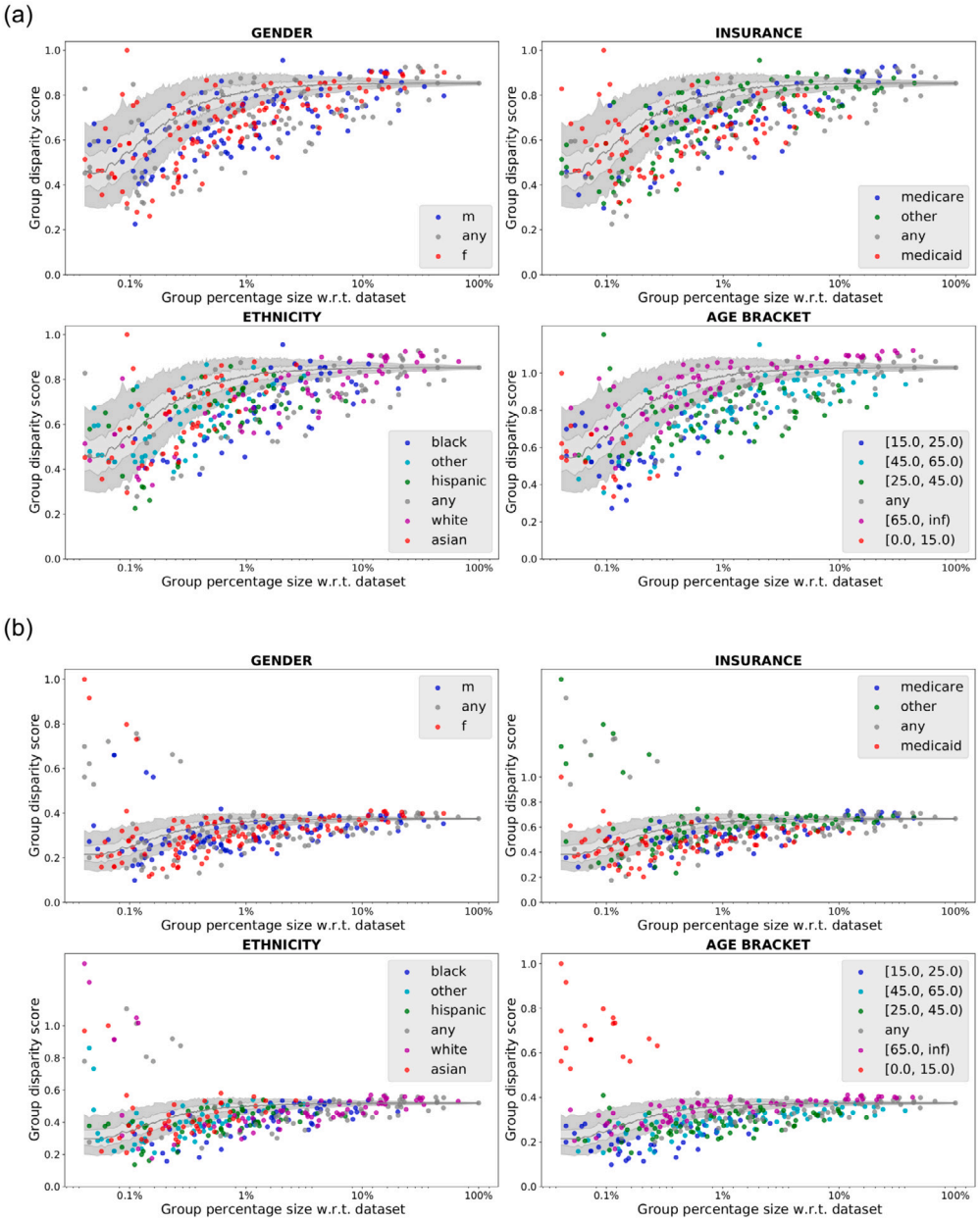

Fig. 7. FairLens scatterplots for the baseline DSS (a) and biased DSS (b).

Table 8
The ranking performed by FairLens using disparity score for the baseline and biased DSS.

Insurance	Rank on baseline	Rank on biased	Mean rank on baseline	Mean rank on biased
medicare	1	2	107.82	119.09
other	2	1	111.24	93.97
medicaid	3	3	143.75	155.04

Finally, we measured the outcome disparity for the insurance variable using the multi-label standard metrics used to evaluate DoctorAI performance in the original paper, *recall@k* and the *microAUC*. We compared the difference of these metrics in the baseline and biased DSS in Fig. 8. We can see that while the standard metrics remain almost constant or slightly decrease in the biased BB

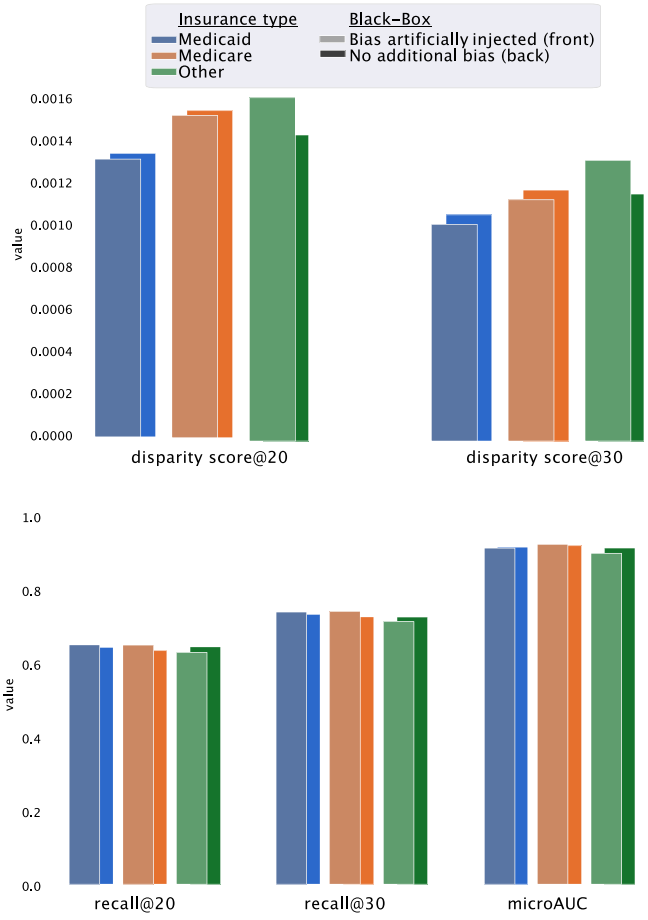

Fig. 8. Average metrics across insurance groups.

with respect to the baseline DSS, both disparity scores evaluated at top $k = 20$ and $k = 30$ exhibit a clear increase for the under sampled group in the BB where bias was artificially injected.

6. Discussion and conclusions

Fairness and explainability are key features to gain trust from patients and clinicians. As black-box ML-based clinical decision support systems will be deployed in real-world healthcare settings, systematic auditing procedures must be in place. In this paper we proposed FairLens, an algorithmic pipeline to inspect clinical DSSs to spot potential fairness issues in patients' groups that call for further investigation of potential over-/under-diagnosed conditions. The proposed methodology is able to drive domain experts to investigate the reason behind the systematic black-box misclassification by pointing to the most common causes of error within groups through XAI techniques.

The main use-case presented in this paper describes the auditing process of a clinical decision support system trained on sequential visits aimed at predicting the diagnoses associated to the next patient's visit. FairLens can be generalized to other use-cases with different DSS tasks, as far as the building blocks are adequately adapted. In the Supplementary Information, the application of FairLens to a different clinical decision support system is shown. While the final aim (auditing a black-box) and the intended user (IT expert responsible for deploying the DSS in the healthcare facility) are the same, the machine learning model is completely different. The experiment highlights the flexibility of our framework, that is adapted to work on the task of predicting the ICD9 codes given the raw text of clinical notes, relaxing the temporal dimension of sequential visits. While the scoring mechanism remains unchanged, the explainability approach and the local-to-global aggregation mechanism are adapted to the prediction task. We highlight that FairLens can also be used by the DSS developers to perform a sanity check of the model and detect and mitigate potential biases before its release. However, this would require the ML engineers to have some knowledge of the medical domain, or to cooperate with medical personnel, to understand if the potential bias signaled by FairLens reflects a real fairness issue.

It is worth stressing that FairLens is not designed to be an automated tool, but rather to help human auditors in identifying groups where fairness issues may arise. Moreover, FairLens is not able to provide the origin of such misbehavior (e.g. eliciting if

the source of bias is in the original training data, or is embedded in the algorithm itself García-Soriano & Bonchi, 2020 or in the prediction task), as it is designed to perform external audit without having access to information about the black-box nor to the original training data.

We firmly believe in the primary urgency of building external algorithmic auditing tools that allow an objective evaluation of the effectiveness and fairness of algorithmic systems. Even though an internal algorithmic auditing process is of pivotal importance to release a product that meets the ethical and reliability standards of whom developed and marketed the product, its cost–benefit analysis might be skewed toward maximizing profit. External auditing tools allow companies to be held accountable to third parties and increase the credibility of the algorithmic pipeline. Independent auditing is also useful to test the model in the actual deployment setting: it may also happen that the population used for training the DSS is just not compatible with the target population where the DSS should be deployed, thus making it harmful or ineffective. In the healthcare context, external auditing tools such as FairLens could also identify ICU patients' over/under-treatment to improve patient-processes. Under the assumption that high disparity scores suggest a mismatch between what the clinical decision support system learned and how the patients were historically treated in the healthcare facility, the auditor might even find biases in the auditing data, that should lead to additional investigation and quality assessment of hospital services. It is also important to discuss potential uses of FairLens, which differ from the one envisioned and discussed in this paper. Theoretically, if linked with information that leads to the identification of the operator responsible for patients' treatments, FairLens could be used to identify doctors that systematically treat groups differently. While doctor performance assessment is extremely valuable and several techniques to operationalize it already exist (Overeem et al., 2007), such unintended use of FairLens should be properly considered.

Future work will be devoted to test FairLens in a setting with a panel of domain experts in the loop to optimize tool usability and understand if the provided metrics, explanations and investigation steps are meaningful and understandable by the end-users, and if FairLens is ultimately helpful in taking better informed decisions on DSS deployment. Finally, additional experiments to generate counterfactual examples as explanations will be implemented to increase FairLens adoption by domain experts.

CRediT authorship contribution statement

Cecilia Panigutti: Conceptualization, Methodology, Software, Writing - original draft, Review & editing. **Alan Perotti:** Conceptualization, Methodology, Software, Writing - original draft, Review & editing. **André Panisson:** Conceptualization, Methodology, Software, Writing - original draft, Review & editing. **Paolo Bajardi:** Conceptualization, Methodology, Writing - original draft, Review & editing. **Dino Pedreschi:** Conceptualization, Methodology, Supervision.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.ipm.2021.102657.

References

Abdollahpouri, H., Burke, R., & Mobasher, B. (2017). Controlling popularity bias in learning-to-rank recommendation. In *Proceedings of the eleventh ACM conference on recommender systems* (pp. 42–46).

Abràmoff, M. D., Lavin, P. T., Birch, M., Shah, N., & Folk, J. C. (2018). Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. *NPJ Digital Medicine*, *1*(1), 1–8.

Adebayo, J. A., et al. (2016). *Fairml: Toolbox for diagnosing bias in predictive modeling* (Ph.D. thesis), Massachusetts Institute of Technology.

Angelino, E., Larus-Stone, N., Alabi, D., Seltzer, M., & Rudin, C. (2017). Learning certifiably optimal rule lists for categorical data. *Journal of Machine Learning Research*, *18*(1), 8753–8830.

Anjomshoae, S., Kampik, T., & Främling, K. (2020). Py-CIU: A python library for explaining machine learning predictions using contextual importance and utility. In *IJCAI-PRICAI 2020 workshop on explainable artificial intelligence (XAI)*.

Avati, A., Jung, K., Harman, S., Downing, L., Ng, A., & Shah, N. H. (2018). Improving palliative care with deep learning. *BMC Medical Informatics and Decision Making*, *18*(4), 122.

Barocas, S., Hardt, M., & Narayanan, A. (2017). Fairness in machine learning. *Nips Tutorial*, *1*, 2.

Bejnordi, B. E., Veta, M., Van Diest, P. J., Van Ginneken, B., Karssemeijer, N., Litjens, G., et al. (2017). Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *Jama*, *318*(22), 2199–2210.

Bellamy, R. K., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., et al. (2018). AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. arxiv preprint arxiv:1810.01943.

Boag, W., Suresh, H., Celi, L. A., Szolovits, P., & Ghassemi, M. (2018). Racial disparities and mistrust in end-of-life care. In *Machine learning for healthcare conference* (pp. 587–602). PMLR.

Capper, D., Jones, D. T., Sill, M., Hovestadt, V., Schrimpf, D., Sturm, D., et al. (2018). DNA Methylation-based classification of central nervous system tumours. *Nature*, *555*(7697), 469–474.

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (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* (pp. 1721–1730).

Casey, J. A., Schwartz, B. S., Stewart, W. F., & Adler, N. E. (2016). Using electronic health records for population health research: a review of methods and applications. *Annual Review of Public Health*, *37*, 61–81.

Che, Z., Purushotham, S., Cho, K., Sontag, D., & Liu, Y. (2018). Recurrent neural networks for multivariate time series with missing values. *Scientific Reports*, *8*(1), 1–12.

Chen, M., Hao, Y., Hwang, K., Wang, L., & Wang, L. (2017). Disease prediction by machine learning over big data from healthcare communities. *Ieee Access*, *5*, 8869–8879.

Chen, I. Y., Pierson, E., Rose, S., Joshi, S., Ferryman, K., & Ghassemi, M. (2020). Ethical machine learning in health. arxiv preprint arxiv:2009.10576.

Chen, I. Y., Szolovits, P., & Ghassemi, M. (2019). Can AI help reduce disparities in general medical and mental health care?. *AMA Journal of Ethics*, *21*(2), 167–179.

Chilamkurthy, S., Ghosh, R., Tanamala, S., Biviji, M., Campeau, N. G., Venugopal, V. K., et al. (2018). Deep learning algorithms for detection of critical findings in head ct scans: a retrospective study. *The Lancet*, *392*(10162), 2388–2396.
Choi, E., Bahadori, M. T., Schuetz, A., Stewart, W. F., & Sun, J. (2016). Doctor AI: Predicting clinical events via recurrent neural networks. In *Machine learning for healthcare conference* (pp. 301–318).
Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, *5*(2), 153–163.
Coudray, N., Ocampo, P. S., Sakellaropoulos, T., Narula, N., Snuderl, M., Fenyö, D., et al. (2018). Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. *Nature Medicine*, *24*(10), 1559–1567.
Cutillo, C. M., Sharma, K. R., Foschini, L., Kundu, S., Mackintosh, M., & Mandl, K. D. (2020). Machine intelligence in healthcare—perspectives on trustworthiness, explainability, usability, and transparency. *NPJ Digital Medicine*, *3*(1), 1–5.
Davenport, T., & Kalakota, R. (2019). The potential for artificial intelligence in healthcare. *Future Healthcare Journal*, *6*(2), 94.
Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference* (pp. 214–226).
Edizel, B., Bonchi, F., Hajian, S., Panisson, A., & Tassa, T. (2020). Fairecsys: Mitigating algorithmic bias in recommender systems. *International Journal of Data Science and Analytics*, *9*(2), 197–213.
Ellis, C., & Jacobs, M. (2021). The complexity of health disparities: More than just black–white differences. *Perspectives of the ASHA Special Interest Groups*, 1–10.
Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., et al. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, *542*(7639), 115–118.
Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 259–268).
Frogner, C., Zhang, C., Mobahi, H., Araya, M., & Poggio, T. A. (2015). Learning with a wasserstein loss. In *Advances in neural information processing systems* (pp. 2053–2061).
García-Soriano, D., & Bonchi, F. (2020). Fair-by-design matching. *Data Mining and Knowledge Discovery*, 1–45.
Geruso, M., & Layton, T. (2020). Upcoding: Evidence from medicare on squishy risk adjustment. *Journal of Political Economy*, *128*(3), 984–1026.
Goddard, K., Roudsari, A., & Wyatt, J. C. (2012). Automation bias: a systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association*, *19*(1), 121–127.
Goldberger, A. L., Amaral, L. A., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., et al. (2000). Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *Circulation*, *101*(23), e215–e220.
Guidotti, R., Monreale, A., Matwin, S., & Pedreschi, D. (2020). Explaining image classifiers generating exemplars and counter-exemplars from latent representations. In *Proceedings of the AAAI conference on artificial intelligence, Vol. 34* (pp. 13665–13668).
Guidotti, R., Monreale, A., Ruggieri, S., Pedreschi, D., Turini, F., & Giannotti, F. (2018). Local rule-based explanations of black box decision systems. arxiv preprint arxiv:1805.10820.
Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys*, *51*(5), 1–42.
Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., 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.
Gunning, D. (2017). *Explainable artificial intelligence (xai), Vol. 2*. Defense Advanced Research Projects Agency (DARPA), Nd Web.
Haenssle, H. A., Fink, C., Schneiderbauer, R., Toberer, F., Buhl, T., Blum, A., et al. (2018). Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Annals of Oncology*, *29*(8), 1836–1842.
Hajian, S., Bonchi, F., & Castillo, C. (2016). Algorithmic bias: From discrimination discovery to fairness-aware data mining. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 2125–2126).
Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. In *Advances in neural information processing systems* (pp. 3315–3323).
Harrington, K., Allen, A., & Ruchala, L. (2007). Restraining medicare abuse: the case of upcoding. *Research in Healthcare Financial Management*, *11*(1), 1.
Heiat, A., Gross, C. P., & Krumholz, H. M. (2002). Representation of the elderly, women, and minorities in heart failure clinical trials. *Archives of Internal Medicine*, *162*(15).
Hillson, S. D., Connelly, D. P., & Liu, Y. (1995). The effects of computer-assisted electrocardiographic interpretation on physicians' diagnostic decisions. *Medical Decision Making*, *15*(2), 107–112.
Jiang, F., Jiang, Y., Zhi, H., Dong, Y., Li, H., Ma, S., et al. (2017). Artificial intelligence in healthcare: past, present and future. *Stroke and Vascular Neurology*, *2*(4), 230–243.
Jiang, R., Pacchiano, A., Stepleton, T., Jiang, H., & Chiappa, S. (2020). Wasserstein fair classification. In *Uncertainty in artificial intelligence* (pp. 862–872). PMLR.
Johnson, A., Bulgarelli, L., Pollard, T., Horng, S., Celi, L. A., & Roger, M. (2020). MIMIC-IV (Version 0.4). *PhysioNet*, http://dx.doi.org/10.13026/a3wn-hq05.
Johnson, A. E., Pollard, T. J., Shen, L., Li-wei, H. L., Feng, M., Ghassemi, M., et al. (2016). MIMIC-III, a freely accessible critical care database. *Scientific Data*, *3*, Article 160035.
Kearns, M., Neel, S., Roth, A., & Wu, Z. S. (2018). Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International conference on machine learning* (pp. 2564–2572).
Keppel, K., Pamuk, E., Lynch, J., Carter-Pokras, O., Kim, I., Mays, V., et al. (2005). Methodological issues in measuring health disparities. *Vital and Health Statistics. Series 2, Data Evaluation and Methods Research*, (141), 1.
Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. arxiv preprint arxiv:1609.05807.
Letham, B., Rudin, C., McCormick, T. H., Madigan, D., et al. (2015). Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model. *Annals of Applied Statistics*, *9*(3), 1350–1371.
Lindow, T., Kron, J., Thulesius, H., Ljungström, E., & Pahlm, O. (2019). Erroneous computer-based interpretations of atrial fibrillation and atrial flutter in a Swedish primary health care setting. *Scandinavian Journal of Primary Health Care*, *37*(4), 426–433.
Lindsey, R., Daluiski, A., Chopra, S., Lachapelle, A., Mozer, M., Sicular, S., et al. (2018). Deep neural network improves fracture detection by clinicians. *Proceedings of the National Academy of Sciences*, *115*(45), 11591–11596.
Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., et al. (2020). From local explanations to global understanding with explainable ai for trees. *Nature Machine Intelligence*, *2*(1), 2522–5839.
Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in neural information processing systems* (pp. 4765–4774).
Luong, B. T., Ruggieri, S., & Turini, F. (2011). k-NN as an implementation of situation testing for discrimination discovery and prevention. In *Proceedings of the 17th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 502–510).
Lyon, S. M., Benson, N. M., Cooke, C. R., Iwashyna, T. J., Ratcliffe, S. J., & Kahn, J. M. (2011). The effect of insurance status on mortality and procedural use in critically ill patients. *American Journal of Respiratory and Critical Care Medicine*, *184*(7), 809–815.
Madani, A., Arnaout, R., Mofrad, M., & Arnaout, R. (2018). Fast and accurate view classification of echocardiograms using deep learning. *NPJ Digital Medicine*, *1*(1), 1–8.
Mason, S., Hussain-Gambles, M., Leese, B., Atkin, K., & Brown, J. (2003). Representation of south Asian people in randomised clinical trials: analysis of trials' data. *Bmj*, *326*(7401), 1244–1245.

McMaughan, D. J., Oloruntoba, O., & Smith, M. L. (2020). Socioeconomic status and access to healthcare: Interrelated drivers for healthy aging. *Frontiers in Public Health*, *8*.

Miranda-Escalada, A., Gonzalez-Agirre, A., Armengol-Estapé, J., & Krallinger, M. (2020). Overview of automatic clinical coding: annotations, guidelines, and solutions for non-english clinical cases at codiesp track of CLEF ehealth 2020. In *Working notes of conference and labs of the evaluation (clef) forum. ceur workshop proceedings*.

Moja, L., Friz, H. P., Capobussi, M., Kwag, K., Banzi, R., Ruggiero, F., et al. (2019). Effectiveness of a hospital-based computerized decision support system on clinician recommendations and patient outcomes: A randomized clinical trial. *JAMA Network Open*, *2*(12), e1917094.

Mullenbach, J., Wiegreffe, S., Duke, J., Sun, J., & Eisenstein, J. (2018). Explainable prediction of medical codes from clinical text. In *Proceedings of the 2018 conference of the north american chapter of the association for computational linguistics: human language technologies, volume 1 (long papers)* (pp. 1101–1111). New Orleans, Louisiana: Association for Computational Linguistics.

Nam, J. G., Park, S., Hwang, E. J., Lee, J. H., Jin, K.-N., Lim, K. Y., et al. (2019). Development and validation of deep learning–based automatic detection algorithm for malignant pulmonary nodules on chest radiographs. *Radiology*, *290*(1), 218–228.

Norgeot, B., Glicksberg, B. S., & Butte, A. J. (2019). A call for deep-learning healthcare. *Nature Medicine*, *25*(1), 14–15.

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

O'malley, K. J., Cook, K. F., Price, M. D., Wildes, K. R., Hurdle, J. F., & Ashton, C. M. (2005). Measuring diagnoses: ICD code accuracy. *Health Services Research*, *40*(5p2), 1620–1639.

Overeem, K., Faber, M. J., Arah, O. A., Elwyn, G., Lombarts, K. M., Wollersheim, H. C., et al. (2007). Doctor performance assessment in daily practise: does it help doctors or not? A systematic review. *Medical Education*, *41*(11), 1039–1049.

Panigutti, C., Guidotti, R., Monreale, A., & Pedreschi, D. (2019). Explaining multi-label black-box classifiers for health applications. In *International workshop on health intelligence* (pp. 97–110). Springer.

Panigutti, C., Perotti, A., & Pedreschi, D. (2020). Doctor XAI: an ontology-based approach to black-box sequential data classification explanations. In *Proceedings of the 2020 conference on fairness, accountability, and transparency* (pp. 629–639).

Pedreschi, D., Ruggieri, S., & Turini, F. (2008). Discrimination-aware data mining. In *Proceedings of the 14th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 560–568).

Pierson, E., Cutler, D. M., Leskovec, J., Mullainathan, S., & Obermeyer, Z. (2021). An algorithmic approach to reducing unexplained pain disparities in underserved populations. *Nature Medicine*, *27*(1), 136–140.

Piper, C. (2013). Popular health care provider fraud schemes. *Association of Certified Fraud Examiners*.

Polignano, M., Suriano, V., Lops, P., de Gemmis, M., & Semeraro, G. (2020). A study of machine learning models for clinical coding of medical reports at codiesp 2020. In *Working notes of conference and labs of the evaluation (clef) forum. ceur workshop proceedings*.

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., et al. (2020). Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 conference on fairness, accountability, and transparency* (pp. 33–44).

Rajkomar, A., Oren, E., Chen, K., Dai, A. M., Hajaj, N., Hardt, M., et al. (2018). Scalable and accurate deep learning with electronic health records. *NPJ Digital Medicine*, *1*(1), 18.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135–1144).

Ribeiro, M. T., Singh, S., & Guestrin, C. (2018). Anchors: High-precision model-agnostic explanations. In *Proceedings of the AAAI conference on artificial intelligence, Vol. 32*.

Ruggieri, S., Pedreschi, D., & Turini, F. (2010). Data mining for discrimination discovery. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, *4*(2), 1–40.

Saleiro, P., Kuester, B., Hinkson, L., London, J., Stevens, A., Anisfeld, A., et al. (2018). Aequitas: A bias and fairness audit toolkit. arxiv preprint arxiv:1811.05577.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision* (pp. 618–626).

Setzu, M., Guidotti, R., Monreale, A., & Turini, F. (2019). Global explanations with local scoring. In *Joint european conference on machine learning and knowledge discovery in databases* (pp. 159–171). Springer.

Setzu, M., Guidotti, R., Monreale, A., Turini, F., Pedreschi, D., & Giannotti, F. (2021). GlocalX-From local to global explanations of black box AI models. *Artificial Intelligence*, Article 103457.

Seyyed-Kalantari, L., Liu, G., McDermott, M., & Ghassemi, M. (2020). Chexclusion: Fairness gaps in deep chest X-ray classifiers. arxiv:2003.00827.

Shameer, K., Johnson, K. W., Yahi, A., Miotto, R., Li, L., Ricks, D., et al. (2017). Predictive modeling of hospital readmission rates using electronic medical record-wide machine learning: a case-study using mount sinai heart failure cohort. In *Pacific symposium on biocomputing 2017* (pp. 276–287). World Scientific.

Titano, J. J., Badgeley, M., Schefflein, J., Pain, M., Su, A., Cai, M., et al. (2018). Automated deep-neural-network surveillance of cranial images for acute neurologic events. *Nature Medicine*, *24*(9), 1337–1341.

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

Tramer, F., Atlidakis, V., Geambasu, R., Hsu, D., Hubaux, J.-P., Humbert, M., et al. (2017). Fairtest: Discovering unwarranted associations in data-driven applications. In *2017 IEEE european symposium on security and privacy (EuroS&P)* (pp. 401–416). IEEE.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. arxiv preprint arxiv:1706.03762.

Wang, T., Rudin, C., Doshi-Velez, F., Liu, Y., Klampfl, E., & MacNeille, P. (2017). A bayesian framework for learning rule sets for interpretable classification. *Journal of Machine Learning Research*, *18*(1), 2357–2393.

WHO, W. H. O., et al. (2018). ICD Purpose and uses. *Classification*, Available Online at: http://www.who.int/classifications/icd/en/ (accessed May 20, 2020).

Wiegreffe, S., & Pinter, Y. (2019). Attention is not not explanation. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)* (pp. 11–20). Hong Kong, China: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/D19-1002, https://www.aclweb.org/anthology/D19-1002.

Xiao, C., Choi, E., & Sun, J. (2018). Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. *Journal of the American Medical Informatics Association*, *25*(10), 1419–1428.

Yu, K.-H., Beam, A. L., & Kohane, I. S. (2018). Artificial intelligence in healthcare. *Nature Biomedical Engineering*, *2*(10), 719–731.

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., & Dwork, C. (2013). Learning fair representations. In S. Dasgupta, & D. McAllester (Eds.), *Proceedings of the 30th international conference on machine learning, Vol. 3* (pp. 325–333). Atlanta, Georgia, USA: PMLR.

Zhang, J., Gajjala, S., Agrawal, P., Tison, G. H., Hallock, L. A., Beussink-Nelson, L., et al. (2018). Fully automated echocardiogram interpretation in clinical practice: feasibility and diagnostic accuracy. *Circulation*, *138*(16), 1623–1635.

Zhang, X., Tan, S., Koch, P., Lou, Y., Chajewska, U., & Caruana, R. (2019). Interpretability is harder in the multiclass setting: axiomatic interpretability for multiclass additive models. *Age*, *25*(50), 75–100.