# InterFair: Intersectional Fairness with Fairness-Oriented Multiobjective Optimization

Willam La Cava & Elle Lett
Team Lead Contact: williamlacava@gmail.com, (413) 320-0544
Video: https://www.youtube.com/watch?v=nf-J-1pqvVk

## Abstract

*InterFair* is a new, highly flexible tool capable of measuring intersectional fairness and optimizing machine learning (ML) models for fair performance across many settings. InterFair is designed for intersectional fairness, capable of measuring and mitigating bias across groups defined by the intersection of multiple sensitive attributes. To mitigate bias, InterFair leverages fairness-oriented multiobjective optimization (FOMO), a genetic algorithm-based approach that iteratively updates a series of model training weights to optimize for pre-specified fairness and accuracy objective functions. FOMO returns a pareto frontier of optimal models for the selected accuracy and fairness metrics. We demonstrate the use of InterFair in the healthcare setting for predicting admissions from the emergency room to minimize post-triage wait-times. We show that InterFair can be used to find a model that maximizes subgroup false negative fairness, preventing biased deprioritization for marginalized groups, and minimizes overall false positive rates, preventing unnecessary allocation of hospital beds. It offers a range of fairness and accuracy measures making it adaptable to any scenario where a prediction model can be built and appropriate trade-offs can be defined. In addition, InterFair has built in procedures for updating and re-training the model to prevent model drift. Built using open-source software, InterFair can be adapted by any healthcare-related entity for their use case.

## GitHub code

The code for our submission is available from https://github.com/cavalab/interfair . The associated webpage https://cavalab.org/interfair contains documentation on use, including demonstrations. Throughout this supporting documentation and particularly in the *Healthcare Scenario* section we include output and visualizations to demo our tool.

## Methodology Overview

Our tool is designed to work with most common definitions of fairness. Our current implementation supports equalized false positive rates (FPR) and false negative rates (FNR) (together referred to as equalized odds), equalized mean squared error (MSE), equalized positivity rates (a.k.a. demographic parity), and equalized calibration (specifically multicalibration). Unlike tools such as AI-Fairness 360 (IBM) and fairlearn (Microsoft), InterFair defines these metrics over rich subgroups in the data, for example by defining these measures as variations of **subgroup fairness (SF)** ($SF = max_{g \in G} p_g |m_g - m|$). Briefly, **subgroup fairness** is the maximum deviation of a model's performance among pre-specified groups ($m_g$) from the overall model performance ($\mathrm{m}$), normalized to the group size ($\mathrm{p}_g$). droups can be flexibly defined by one or more sensitive attributes and SF can be used with any accuracy metric. We selected these different metrics because no single metric can be appropriate for all healthcare scenarios. This way, users can specify the metric most appropriate for their use-case.

*Figure 1: Fairness-Oriented Multiobjective Optimization (FOMO) Pipeline*

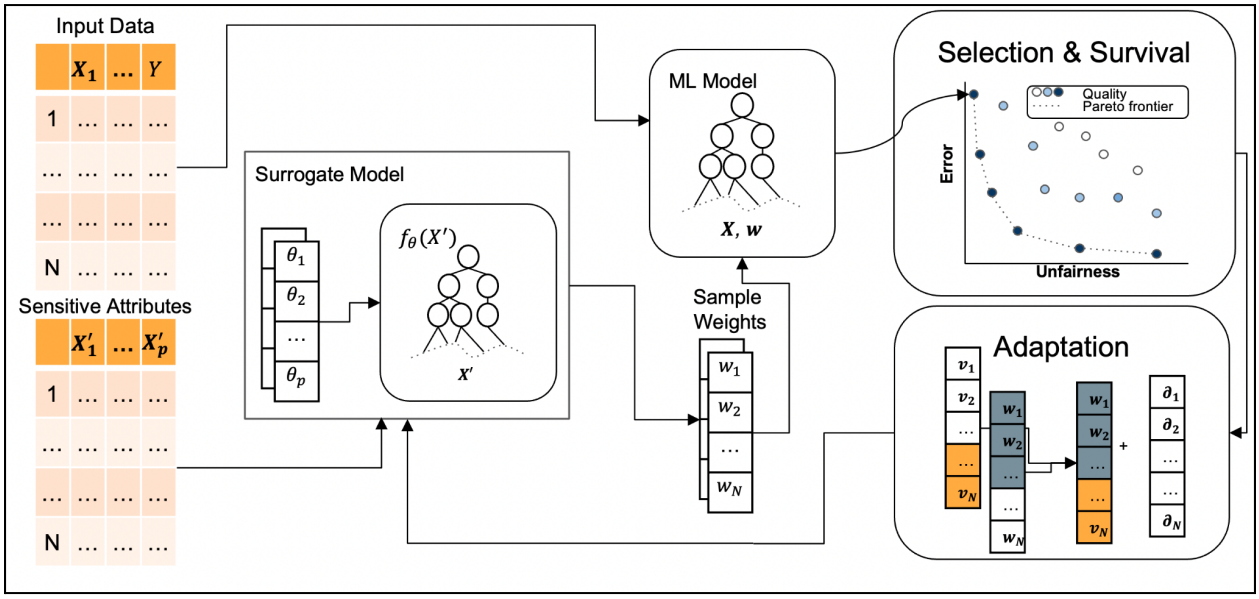



Figure 1 shows the pipeline for FOMO, an open-source bias mitigation tool used by InterFair. FOMO is an evolutionary algorithm that, at a high-level, mirrors the natural selection process that 'optimizes' the genetic material inherited between generations of mating and reproduction. Here, each generation consists of multiple sets of parameters.

These low-dimensional parameters are used by a surrogate model to map the sensitive attributes to the weights in a higher dimension parameter space, thereby reducing computation time.  Then, each set of weights is fed into an ML model, which is, in turn, assessed based on its fairness and error. Importantly, there is an inherent trade-off between fairness and statistical error, or, social and predictive bias, and among all models in the generation there will be variation on where a model falls along the “unfairness” vs error curve. Using an open-source python package for multi-objective optimization, *pymoo*, FOMO selects weights corresponding to models along the **pareto frontier** to ‘survive’ until the next generations. These models represent those in that generation that maximize the fairness-error trade off and move to the adaptation stage where crossover occurs and random noise is added to the weights to mimic mating and random mutation during evolution. The process repeats, selecting for weights that yield ML models that are more fair with less overall error until a stopping criteria on the number of iterations or change in the pareto frontier is reached. This can be adapted for various fairness metrics, as in our example subgroup fairness, and different measures of overall error.

The aforedescribed fitting procedure readily handles retrospective bias. However, InterFair is also well-suited for prospective bias detection. One of the innovations of InterFair is that it saves the training state after a model is selected. This means that, should re-training need to occur, users can “pick-up” at the last iteration of the optimization procedure and, with additional data, update the model. This can allow for a surveillance system that protects against model drift. This is further elaborate in the *Sustainability Plan* section.

## Value Proposition

The primary innovation and value of our tool to healthcare is its flexibility. By using InterFair, we can identify bias that impacts any group defined by one or more sensitive attributes. One of the challenges of fair ML is fairness gerrymandering, wherein, a model can perform “fairly” across groups defined by separate attributes such as race (e.g. similar performance for Black and White patients), and gender (e.g. similar performance for male and female patients ), but not perform well for their intersections (e.g. poor performance for Black females compared to White males). This invokes intersectionality, the social theory and justice framework that explains how individuals who have multiple marginalized identities and are subject to multiple forms of discrimination which can lead to unique  Our tool has the capacity to detect intersectional social biases, thereby identifying  (example from output).

In addition to its ability to detect intersectional social biases, InterFair also allows the users to select amongst different metrics and adapt to the user’s specific prediction task. For example, consider an STI screening scenario in a clinic where the goal is to identify patients at high-risk for gonorrhea or chlamydia. In this case, treatment is low-risk for the patients, and identifying individuals to screen has high community and individual benefits (preventing transmission and negative side effects of prolonged infection). There, we would want to prevent false negatives at the expense of an increase in false-positives and overall error. Further, we would want to use false negatives as the fairness metric to ensure that no one group was being inappropriately disadvantaged by the model and undertested. Conversely, for an ML model to identify patients who were likely to misuse prescription opioid medications, false positives could potentially lead to unjust denial of treatment to patients with severe pain management needs. InterFair would be suitable for both scenarios; the only requirements would be to specify the desired metric (FP or FN) ,provide demographic data on sensitive attributes, and the feature data or predictions from an ML model and our tool can be used to estimate subgroup fairness and update the model for optimized fair performance within a pre-specified constraint. Lastly, the tool will output which group is subject to the greatest fairness violation, thereby identifying which social bias(es) might contribute to inequitable model performance.

InterFair also provides overall error metrics to detect predictive bias and allows selection of models that balances the tradeoff between social and predictive biases as discussed in the *Methodology Overview* section.

## Healthcare Scenario

For the purpose of our submission we demonstrate our tool in the context of predicting admission from the emergency rooms. Emergency rooms represent the first-line of care for patients with varying levels of acuity in a dynamic treatment setting. One of the challenges is bed coordination; identifying patients who will need extended stays for treatment and assigning them hospital beds. Beds are a finite resource and many hospitals operate at or over capacity and must temporarily bed patients in non-treatment spaces (hallways, etc), or subject them to extensive wait times before receiving treatment. These challenges reduce the efficiency of healthcare administration making it difficult to coordinate care and potentially causing patient health status to worsen due to delayed treatment.

In our demo, InterFair helps equitably increase the bed coordination process by optimizing an ML model predicting admission from the emergency room for fairness. We demonstrate our tool on the Medical Information Mart for Intensive Care-IV (MIMIC-IV) data, a publicly available electronic health record database on patients seen in an emergency department at Beth Israel Deaconess Hospital system. The data available in MIMIV-IV includes features; patient vital signs on admission, demographics, and other patient characteristics, and the label, admission status. We assume a scenario where a hospital is operating at or over bed capacity, and that the sensitive attributes of interest are race/ethnicity (referred to as ethnicity in the output figures below), sex, and insurance status. For this scenario, we want fair performance across groups as measured by the false negative rate (FNR) Using the MIMIC-IV data we built a random forest prediction model for admission and the *measure_disparity.py* script to measure overall performance. Below are relevant excerpts from the demo.

*Figure 2: Predictive Bias Measures with InterFair*

```
from measure_disparity import measure_disparity
df_fairness = measure_disparity('base_model_mimic4_admission.csv')
```

```
reading in base_model_mimic4_admission.csv
demographic columns: ['insurance', 'ethnicity', 'gender']
======================================
Overall Performance
======================================
        Measures of predictive bias on the whole population.
```

| AUROC | AUPRC | FPR | FNR | Positivity Rate | Accuracy |
|---|---|---|---|---|---|
| 0.881 | 0.77 | 0.186 | 0.432 | 0.299 | 0.823 |

Figure 2, displays the predictive bias metrics for the base model. By default, it displays the area under receiver operator curve (AUROC), and area under the precision-recall curve (AUPRC) as measures of discrimination, as well as the false positive and false negative rates, overall positivity rate, and accuracy (proportion of correctly classified cases) in the test data.

The base model is reasonably well performing but Figure 3 shows the capacity of InterFair to detect social bias, where the marginal bias metrics (metrics based on performance across groups defined by a single sensitive attribute) and intersectional bias metrics (metrics based on performance across groups defined by multiple sensitive attributes).

*Figure 3: Marginal and Intersectional Fairness Deviations with InterFair*

```
======================================
Subgroup Fairness Violations
======================================
       Measures the deviation in performance for marginal and intersectional groups.
       Note that these deviation are weighted by group prevalence to produce stable estimates when sample sizes are small.
```

| insurance | ethnicity | gender | Brier Score (MSE) | FNR | FPR | Positivity Rate |
|---|---|---|---|---|---|---|
| any | any | F | -0.006 | 0.017 | -0.011 | -0.022 |
| any | any | M | 0.006 | -0.012 | 0.013 | 0.022 |
| any | AMERICAN INDIAN/ALASKA NATIVE | any | 0.0 | 0.0 | -0.0 | -0.0 |
| any | ASIAN | any | 0.001 | 0.002 | 0.0 | 0.0 |
| any | BLACK/AFRICAN AMERICAN | any | -0.012 | !!0.029 | -0.018 | -0.035 |
| any | HISPANIC/LATINO | any | -0.004 | 0.01 | -0.005 | -0.012 |
| any | WHITE | any | !!0.015 | -0.016 | !!0.032 | !!0.047 |
| Medicaid | any | any | -0.005 | 0.006 | -0.007 | -0.013 |
| Medicaid | AMERICAN INDIAN/ALASKA NATIVE | F | -0.0 | 0.0 | -0.0 | -0.0 |
| Medicaid | AMERICAN INDIAN/ALASKA NATIVE | M | -0.0 | -0.0 | -0.0 | -0.0 |
| Medicaid | ASIAN | F | 0.0 | 0.001 | 0.0 | -0.0 |
| Medicaid | ASIAN | M | 0.0 | -0.0 | 0.0 | 0.0 |
| Medicaid | BLACK/AFRICAN AMERICAN | F | -0.002 | 0.005 | -0.003 | -0.006 |
| Medicaid | BLACK/AFRICAN AMERICAN | M | -0.001 | 0.002 | -0.001 | -0.003 |
| Medicaid | HISPANIC/LATINO | F | -0.001 | 0.002 | -0.001 | -0.003 |
| Medicaid | HISPANIC/LATINO | M | -0.0 | 0.001 | -0.0 | -0.001 |
| Medicaid | WHITE | F | -0.0 | 0.001 | -0.0 | -0.001 |
| Medicaid | WHITE | M | -0.0 | -0.001 | -0.0 | -0.0 |
| Medicare | any | any | 0.005 | -0.019 | 0.015 | 0.028 |
| Medicare | AMERICAN INDIAN/ALASKA NATIVE | F | 0.0 | 0.0 | 0.0 | 0.0 |
| Medicare | AMERICAN INDIAN/ALASKA NATIVE | M | 0.0 | 0.0 | 0.0 | 0.0 |
| Medicare | ASIAN | F | 0.0 | 0.0 | 0.0 | 0.0 |
| Medicare | ASIAN | M | 0.0 | -0.0 | 0.0 | 0.0 |
| Medicare | BLACK/AFRICAN AMERICAN | F | -0.001 | 0.003 | -0.002 | -0.003 |
| Medicare | BLACK/AFRICAN AMERICAN | M | -0.0 | 0.001 | -0.001 | -0.002 |
| Medicare | HISPANIC/LATINO | F | -0.0 | 0.0 | -0.0 | -0.001 |
| Medicare | HISPANIC/LATINO | M | -0.0 | 0.0 | 0.0 | -0.0 |
| Medicare | WHITE | F | 0.003 | -0.007 | 0.009 | 0.015 |

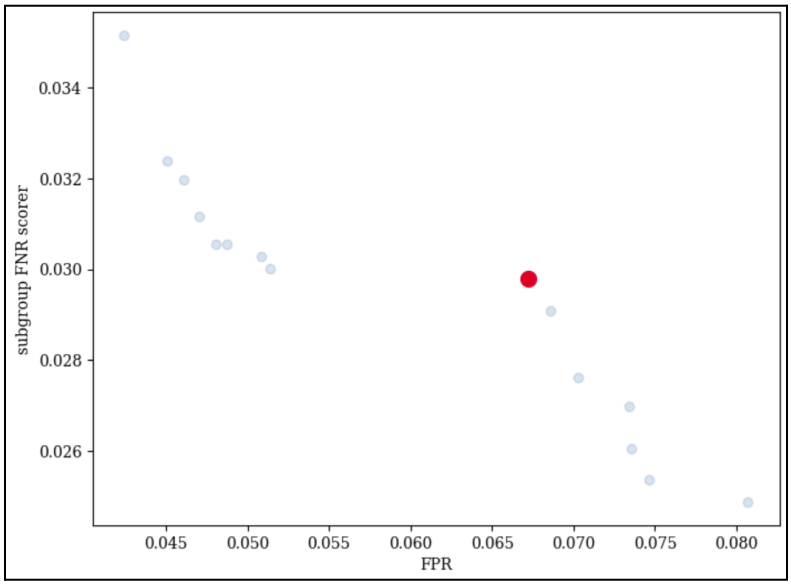
The exclamation points flag the groups with the greatest violations and Figure 4 shows an excerpt of the output that summarizes marginal or intersectional groups with the largest deviations across various subgroup fairness metrics.

*Figure 4: Identification of Groups with Largest Deviations with InterFair*

```
Subgroups with Largest Deviations
---------------------

FNR
----------
- Subgroup: ethnicity=BLACK/AFRICAN AMERICAN
- FNR is 25.5 % higher among this group than the population.

FPR
----------
- Subgroup: ethnicity=WHITE
- FPR is 28.7 % higher among this group than the population.

Positivity Rate
----------
- Subgroup: ethnicity=WHITE
- Positivity Rate is 26.2 % higher among this group than the population.
```

Figure 4 output suggests that Black/African American patients are disproportionately deprioritized by the base model given the high subgroup FNR deviation, suggesting that de-biasing is a necessary intervention to ensure that our model doesn't propagate or exacerbate health inequities.

*Figure 5: Pareto Frontier with InterFair*




Using the *mitigate_bias.py* script, the base model can be optimized for fairness and users can visualize the pareto front as shown in Figure 5. FOMO contains commands to automatically select among the various models the best compromise between fairness and error, for our model subgroup FNR and overall FPR, or the user can select their model directly based on visual inspection. Based on the selected model, shown in red, Figure 6 demonstrates the fairness gains made from applying FOMO to the model.

*Figure 6: Model Improvements from mitigate_bias.py*

| Max Subgroup Deviation in Metric (%) | Original | New |
|---|---|---|
| Brier Score (MSE) | 19.9 | 19.3 |
| Subgroup FNR | 20.4 | 10.9 |
| Subgroup FPR | 86 | 62.3 |
| Positivity Rate | 44.9 | 28.8 |

This table shows the maximum percent deviation across all subgroups defined by the intersection of race, gender, and insurance status, relative to the overall population in the test data. subgroup Brier score, subgroup FNR, and subgroup FPR were all improved by FOMO. While subgroup FNR was the primary goal fairness metric to optimize, FOMO yielded considerable gains in subgroup FPR as well.

As detailed in the *Sustainability Section*, InterFair can be used to detect model drift by re-training the model from the stored training state. Refer to that section for more details.

## Operational Requirements (less than 1 page)

InterFair requires an installation of Python > 3.8, as well as a set of dependencies specified in requirements.txt. We provide an Anaconda environment as well. InterFair does not require a GPU to use, but does benefit from multi-CPU architectures as the algorithm is highly parallelizable. The minimum suggested requirements for use are a four-core machine with at least 8 GB of RAM. It is available under a BSD 3 license.

## Sustainability Plan

As shown in the *Implementation Requirements*, the stakeholders for implementation of this tool is a multidisciplinary team of implementation scientists, clinicians, ethical advisors, IT professionals, and data scientists and hospital administrators. Of particular relevance to sustainability are the ethical advisory board, clinicians, and IT staff. These groups would be able to collaborate to establish criteria on the severity of fairness violations resulting from model drift that would trigger re-training the model through the process described below. Once the initial IT infrastructure described in the *Implementation Requirements* section is developed to integrate model output into EHR, re-training would be very cost-efficient only requiring convening of a subset of stakeholders to approve the new model.

One of the innovations of InterFair is that it saves the training state each iteration. This means that, should re-training need to occur, users can “pick-up” at the last iteration of the optimization procedure and, with additional data, update the model. This could be used as part of a surveillance system that protects against model drift. As time passes from deployment, at pre-planned checkpoints (quarterly or bi-annually, the model can be re-assessed for fairness violations using *measures_disparities.py* and then further optimized with *mitigate_disparities.py* picking up at the last iteration such that InterFair will provide consistently fair models in response to a dynamic clinical environment. This allows for the model to identify retrospective and prospective predictive bias (changes in overall accuracy) and social bias (changes in fairness). Additionally, because *measure_disparities.py* outputs the marginal (along groups defined by single sensitive attributes) and intersectional group performance metrics, InterFair can help users identify which group(s) experience the most severe social bias.

## Generalizability Plan

InterFair is highly generalizable across three dimensions: 1) the types of ML models that can be used, 2) the fairness metrics that can be measured and optimized 3) and the optimization algorithms that can be applied to select amongst possible candidates in the iterative updating approach that selects the pareto frontier of error vs fairness trade-offs (see *Methodology Overview* for details). InterFair can be applied to any ML algorithm that relies on sample weights in the training process which includes all ML models commonly used in healthcare settings: linear and logistic regression, support vector machines, decision trees, random forests, gradient boosted trees, and neural networks among others. The only additional requirement for our implementation is that the model be scikit-learn compatible. This requirement is not a large barrier because most commonly used ML models are already available in scikit-learn, and the ubiquity of the package makes it a primary tool for development of novel models. Depending on the use case, electronic health record (EHR) front-end development may be necessary to display the output from our tool to clinicians to use during their decision making process (see *Implementation Requirements*).

Additionally, all software dependencies used in this implementation are open source. Therefore, healthcare entities can adopt InterFair for their own practices without cost. This will provide rapid benefits for patients across many settings because our technology is not restricted to resource-rich academic medical systems or private health systems. Any entity that can assemble the appropriate human capital and expertise can implement InterFair without the additional cost burden of expensive software. Therefore, our tool can readily generalize to any healthcare specialty and region.  This also has implications for research and scalability; as the tool is optimized for different use scenarios it is readily transportable to quickly scale up its use,

facilitated high-quality, large-sample, regionally diverse studies of the impact of fair ML models for a given scenario.

Lastly, InterFair's bias mitigation uses FOMO which is built on *pymoo* (python multi-objective optimization), an open source multi-objective optimization genetic algorithm package. Our current implementation uses the Non-dominated Sorted Genetic Algorithm-II (NSGA-II), but other genetic algorithms can easily be substituted with different statistical properties for defining the Pareto front.

As is, InterFair has broad applicability and can be implemented in many healthcare settings with appropriate front-end integration in the EHR.These settings span any type of clinical decision for which a prediction task can be envisioned including, prognostication, diagnosis, and treatment response, and across all clinical specialties (cardiology, oncology, etc). Examples include optimizing fair calibration across racial groups and cancer subtypes using the multicalibration fairness metric when predicting tumor response to a new chemo regimen in oncology  Future steps would be to expand the fairness metrics that are currently available in the package.

## Implementation Requirements

Given the high degree of flexibility and generalizability of InterFair it would be challenging to identify an implementation strategy that is completely inclusive of all the possible use cases of our tool. Instead, we will include a general framework and demonstrate it in the context of the task we used for demonstration, predicting admission from the emergency department. Our implementation strategy presumes an existing set of identified model features relevant to the prediction task. Figure 7 shows the time course of our implementation strategy.

*Figure 7: Implementation Strategy*

| | | | | | |
|---|---|---|---|---|---|
| Step 1: Identify Stakeholders and Sensitive Attributes | | | | | |
| Step 2: Select Metrics, Measure and Mitigate Bias with FOMO | | | | | |
| Step 3: Model Deployment | | | | | |
| Step 4: Post-Deployment Impact Studies | | | | | |
| Step 5 Model Surveillance and Updating With FOMO | | | | | |

Step 1: Identify Stakeholders and Sensitive Attributes

The first step is to identify stakeholders appropriate for the use case. For our prediction task, these stakeholders include all healthcare personnel involved in the bed assignment process for ER admissions, as well as IT professionals who develop the electronic health record (EHR) and how information is displayed on the front-end of the EHR, implementation scientists who can design the process for deploying the tool and evaluating its efficacy, an ethical advisory board to help determine guidelines for the maximum admissible fairness violation, and administration to help advocate for the overall needs of the hospital. For the specific task of predicting admissions from the emergency room, this team of stakeholders would collaborate to identify the populations most at risk for harm from the ML model, which would include marginalized groups that disproportionately appear in the emergency room with high-acuity presentations. For the purpose of our demo, we assumed these were racial/ethnic minority groups from low-income backgrounds. This mapped onto sensitive attributes of race and insurance status (as a proxy for

identifying people who are low-income as those with Medicare/Medicaid, and others on private insurance), and gender to allow InterFair to correct for any bias along that identity.

*Table 1: Stakeholders for InterFair Implementation of Fair ML Model to Predict ER Admissions*

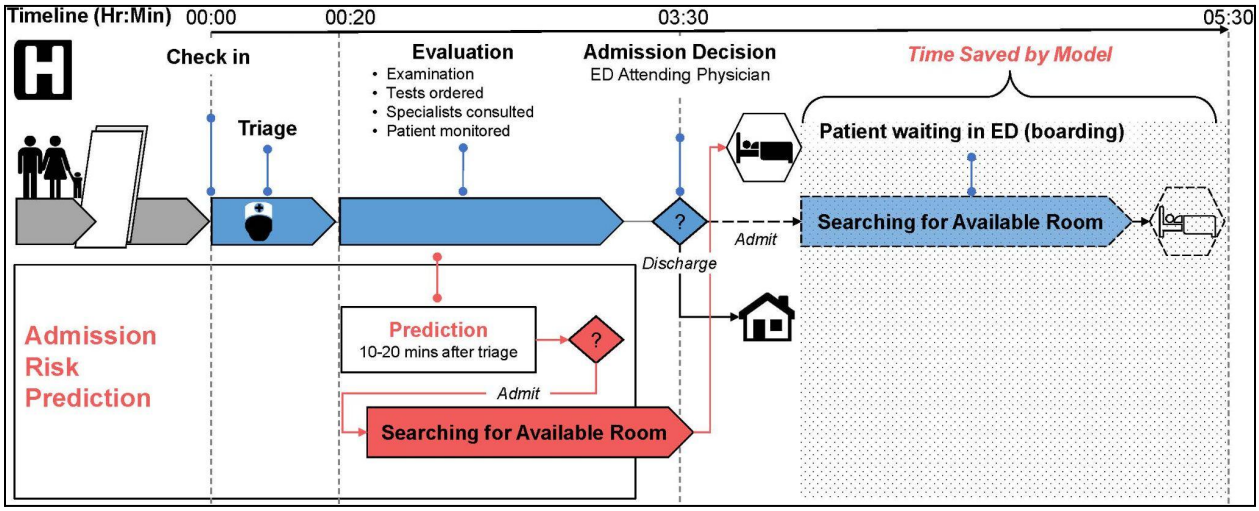
| Stakeholder Group | Members |
|---|---|
| Healthcare Professionals | Emergency Room Clinician and Staff (Nurses, Physicians, Physician Associates, Bed Coordination Team) |
| Healthcare IT | Data Scientists and ML Engineers, Electronic Health Record Designers |
| Ethical Advisory Board | Community Members, Bioethicists, Subject Matter Experts (Computer Scientists, Informaticians and Sociologists) |
| Quality Improvement and Implementation Experts | Implementation Scientists |
| Administration | Hospital C-Suite (CMO, CFO, CEO, etc) |

Step 2: Select Fairness Metric, Accuracy Measure,  Measure Bias, and Mitigate Bias

In step 2 the stakeholders identify the appropriate fairness metric for the use case. In our demo, we assume a hospital system that is often close to or exceeding their bed capacity. In disbursing this finite resource, as described in the *Healthcare Scenario* section, we want to minimize bias in the false negative rate (FNR) so that individuals from specific race-gender-insurance status groups are not unfairly deprioritized for bed space, while minimizing the overall false positive rate (FPR) to ensure we are not allocating precious bed space for individuals who do not need them. With this fairness metric (FNR) and accuracy measure (FPR), InterFair can be used to measure the bias in a base model that is not optimized for fairness to mitigate bias. InterFair will provide the accuracy and fairness metric for the overall data, by groups across each sensitive attribute, and by the intersections of those attributes. This will allow users to identify differential performance by group at baseline. InterFair will also highlight groups with the greatest fairness violations (as shown in the *Healthcare Scenario*) section. Then, the FOMO fitting procedure can be applied to visualize the pareto front (see *Methodology Overview* for details) of possible fairness-aware models that optimize the subgroup fairness for the FNR for a given overall FPR. Then, the stakeholders can select a model that provides them the appropriate trade off (e.g, the hospital system can only “afford” a 5% FPR) and use that model in the deployment phase.

Step 3: Deploy Model

*Figure 8: Model Integration into Triage and Admission Process*




Deploying the model selected in the previous stage will be a collaboration led by the implementation scientists (IS), frontline ER clinical staff, and healthcare IT. IT professionals who manage the front-end of the hospital electronic health record can build data fields that display, in real-time, the model prediction for a patient as clinical data on vital signs and patient characteristics are input into the EHR during patient intake and triage. The IS stakeholders will conduct focus groups of clinical staff and to identify the optimal presentation of the prediction. Prior to, during, or after the patient visit? Binary prediction of “Yes, patient will need a bed”, or continuous probabilities? These preferences can be based on focus group feedback and evaluated in the next step. Figure 3 shows a hypothetical implementation of a FOMO-optimized prediction model for ER admissions and how it would reduce wait times for admitted patients.

Step 4: Post-Deployment Impact Study

After deployment, the team of stakeholders should measure the impact of implementing the model. In the bed coordination task, stakeholders should assess model calibration (were patients of a predicted admission probability actually admitted with that frequency). Based on the fairness metric and sensitive attributes selected in step 1, stakeholders should also assess the differential performance across groups to verify they are within pre-specified admissible boundaries. Beyond the model performance, the stakeholders should evaluate process elements, including a qualitative study on how clinicians using the tool incorporate the model output into their decision making process. Do they immediately call the bed coordination team to see if there is availability and reserve it for their patient when they open the patient chart? Or, do clinicians wait until after their initial assessment of the patient and decide using the model prediction as additional prior information? Does it vary by chief complaint ? These are all questions that will help assess the impact of the tool, and may contribute to how equitably the benefits of the tool are distributed.  Lastly, impact evaluation should extend beyond model performance measures. In our use case, stakeholders should also measure changes in average wait-times for patients, overall, and by intersectional groups. If the model is having a meaningful impact, then the wait-times should decrease, and if it is fair, that decrease should be approximately equivalent across groups. Further, if there is actually greater benefit (more

decrease in wait-times) for the most disadvantaged groups, then the tool will actually be contributing to reducing health inequities.

Step 5: Model Surveillance and Updating

Concurrent with deployment, data from new patients should be continuously added to re-train the model and quarterly stakeholder convenings will allow for selection of optimal models from the updated pareto frontier generated in FOMO. Re-training can occur continuously in the "background" and severe fairness violations can trigger ad-hoc stakeholder meetings. This will ensure that the model continuously performs fairly throughout its use. If, in step 4, new types of biases are identified, new sensitive attributes and fairness metrics can be selected to adapt to the needs of a dynamic healthcare system.

## Lessons Learned

Our tool will allow for assessment of model fairness prior to deployment and continued surveillance across any setting and use case. On the technical side, InterFair only requires that the ML model used for the prediction task be one that uses sample weights in the fitting procedure and be integrated in scikit-learn. With a wide variety of fairness and accuracy metrics InterFair has broad applicability and can be adapted to any healthcare scenario appropriate for ML modeling. It is also very user-friendly allowing for simple visualization of trade-offs between fairness and error making it accessible for all stakeholders that are necessary for true, patient-centered clinical decision making and tool integration. Because of these features, we envision rapid uptake of InterFair across research, academic medical center, and community hospital settings, allowing for robust integration across sectors and high-quality post-impact evaluations. As InterFair becomes ubiquitous, we hope that our fairness-centered tool will engender more trust in AI/ML by showing that a priority of enthusiasm to implement ML-tools in clinical decision making is tempered by doing so in a just way.

One of the challenges we faced in completing this competition was interpreting the instructions. We presume they were purposely spare to allow for maximal flexibility and allowing varying levels of expertise to compete. We solved this challenge by building a package that allows substantial user specification to allow for customized output that can fit the judges', and eventual users' needs.

The data specifications provided for the challenge did not completely map onto real-world manifestations of algorithmic bias. Notably, only including binary sensitive attributes is an oversimplification that will limit the fairness gains from tools like InterFair. For example, racial discrimination and bias, even in data and models, is not simply between White and 'non-White' groups, but often exhibits substantial heterogeneity across racial groups such as Black, Native American, LatinX, and Asian groups, such that binary traits will obscure differences in fairness violations across many of these groups. Without measuring the distinctions in these biases, we cannot mitigate them. Fortunately, InterFair is already capable of handling such data and future competitions on this subject should challenge competitors to build tools that rise to the occasion of mitigating more realistic bias.