# Supplemental Material

## I. Case Studies

### A. Loan Lending Case Study

Given the personal records of individuals such as credit score and income, the financial institution decides to approve or reject the loan application. The predictions of ML model, coupled with the decision policy of the bank, can create a feedback loop that can cause long-term unfairness. Below, we provide the details of the case study.

- **ML model:** We obtained the two predictive models used by D'Amour et al. [1]: 1) max-util agent that aims to maximize the profit (chance of repaying) of the bank, 2) Eq-op agent that myopically maximizes the profit while ensuring the fairness metric measure (Equal Opportunity) between the groups, at each step.
- **Decision maker/policy:** The decision threshold on the probability of repayment is used to decide the approval or rejection of the loan. Each agent employs a different decision policy. The max-util agent computes a static threshold that is supposed to maximize the profit, which is the chance of repayment. The Eq-op agent computes the threshold dynamically to ensure the Equal Opportunity while maximizing the profit. The parameter *bank utility param* is one parameter used to compute the threshold.
- **Environment:** The environment is represented as a dataset containing the credit scores of all the potential loan applicants. We obtained the FICO score dataset, which is based on the 301,536 credit scores preprocessed by [2], and further leveraged by many prior works [1, 3].
- **Dynamics:** The credit score of the individuals are impacted by the system decisions in the following ways:
  1) A loan repayment will increase the individual's credit score by a value sampled from a normal distribution whose mean is equal to the configuration parameter *Score update–repay*. The standard deviation of the normal distribution is determined by the parameter *shift function mode* – "Expectation" has a std of 0, "Normal" has a small std, and "Aggressive" has a large std.
  2) A loan default will reduce the individual's credit score by a value sampled from a normal distribution whose mean is equal to the configuration parameter "Score update–default". Similarly, the std of the normal distribution is determined by the parameter *shift function mode*.
  3) For the projection function, loan applicants are also sampled using a Normal distribution with a mean of 5% and standard deviation of 1% for 20 steps.
- **Results:** The complete regression results for the loan lending case study are shown in Table I. The exact configurations used for the tradeoff analysis in the radar plot of the paper are as follows:
  – Configuration a1: *Score update-repay* = 20, *score update-default* = -40, *shift function mode* = expectation, *agent* = max-util, *bank utility param* = -3.
  – Configuration a2: *Score update-repay* = 8, *score update-default* = -40, *shift function mode* = normal, *agent* = max-util, *bank utility param* = -4.
  – Configuration a3: *Score update-repay* = 8, *score update-default* = -16, *shift function mode* = normal, *agent* = max-util, *bank utility param* = -3.

### B. Opioid Risk Scoring Case Study

Prescription opioid abuse is the leading cause of death in the United States for adults under the age of 50 [4], even more than car crashes or gun violence [5, 6]. Studies have shown that 79.9% of opioid abusers had an opioid prescription before their first abuse [6]. Given the medical records of patients from different groups, the healthcare providers decide on the modification of opioid prescription to reduce abuse risk. However, the modification in the opioid prescription can cause patients to suffer physical and mental debilitation. In this paper, we investigated the dosing discrimination that is induced by the ML-enabled risk-scoring system in the long term.

- **ML model:** Vunikili et al. implemented XGBoost model to predict opioid risk score. We applied the same model structure with one additional input feature – the total number of hospital visits of the patients, which can be changed in the long term. Along with the XGBoost model, we also added a deep neural network model (fully connected network with 8 layers of 64 neurons) for the task with the same input features. As Vunikili et al. used the publicly available Mimic-III dataset for ML model training and evaluation, we used the latest Mimic-IV v2.2 dataset [8]. The dataset consists of more than 15,000,000 medical prescription records for around 300,000 patients obtained from the Beth Israel Deaconess Medical Center in Boston from 2008 to 2019 [7, 9]. From all the records, 80% of patients' information is used for ML model training and evaluation, while the other 20% is used to simulate the environment.
- **Decision maker/policy:** A decision threshold ($[0.3, 0.7]$) is used to decide if the doctor thinks a patient is under opioids abuse risk. If a patient's risk score is greater than the doctor threshold, it is identified as risky and the patient has a chance of multiple hospital visits to get more opioid prescriptions. The exact number of hospital visits and the prescribed amounts of each opioid are affected by the shift function mode.

TABLE I: The complete coefficients and sum of squares of loan lending case study.

| **Terms** | **Dummies** | **Coeffs** | **Sum Sq.** | $\eta^2$**(%)** |
|---|---|---|---|---|
| (Intercept) | | 0.025088*** | 4.71E-03 | 3.03 |
| *Score update–repay* | | 0.000474** | 5.59E-05** | 0.04 |
| *Score update–default* | | -0.003382*** | 2.16E-03*** | 1.39 |
| *Shift function* | aggressive<br>normal | -4.640E-04<br>-1.500E-04 | 7.26E-06 | 0.00 |
| *Agent* | max-util | -0.025079*** | 1.19E-01*** | 76.52 |
| *Bank utility param* | | 0.008445*** | 1.35E-02*** | 8.72 |
| *(Score update–repay, score update–default)* | | 1.570E-04 | 1.89E-05 | 0.01 |
| *(Score update–repay, shift function)* | aggressive<br>normal | 9.100E-05<br>1.060E-04 | 1.68E-06 | 0.00 |
| *(Score update–repay, agent)* | max-util | -0.000539** | 5.58E-05** | 0.04 |
| *(Score update–repay, bank utility param)* | | 0.000327*** | 8.23E-05*** | 0.05 |
| *(Score update–default, shift function)* | aggressive<br>normal | 7.600E-05<br>1.200E-05 | 8.51E-07 | 0.00 |
| *(Score update–default, agent)* | max-util | 0.003354*** | 2.16E-03*** | 1.39 |
| *(Score update–default, bank utility param)* | | -0.000469*** | 1.69E-04*** | 0.11 |
| *(Shift function, agent)* | (aggressive, max-util)<br>(normal, max-util) | 4.620E-04<br>1.480E-04 | 7.11E-06 | 0.00 |
| *(Shift function, bank utility param)* | aggressive<br>normal | -1.150E-04<br>-5.200E-05 | 1.70E-06 | 0.00 |
| *(Agent, bank utility param)* | max-util | -0.008382*** | 1.35E-02 | 8.69 |

***p $<$ .001; **p $<$ .01; *p $<$ .05

- **Environment:** The environment is represented as a dataset containing medical records for patients. Specifically, 20% of patients' information from Mimic-IV v2.2 dataset is used to represent the environment. The medical record for each patient includes a total of 26 features which are used for ML model to make predictions.
- **Dynamics:** The ML prediction brings two changes in the environment: (1) the total number of hospital visits the patients have and (2) the total amount of prescriptions the patients received for each opioid. Two respective shift functions are defined for them.
  - *Shift function for updating hospital visits:* When a patient is evaluated as not under risk, their total number of hospital visits to get opioid prescription will increment by 1. When they are identified under risk, an expected number of hospital visits needed will be computed based on their risk score. The higher the risk score, the greater number of hospital visits possible. After computing the expected number, the exact number of hospital visits will be sampled from a normal distribution whose mean is equal to the pre-computed expected number. The standard deviation of the normal distribution is determined by the parameter, *shift function mode–hospital*. "Expectation" has a std of 0, "Normal" has a small std, and "Aggressive" has a large std. One exception is mode "Equal", which would always treat patients as not risky and update patient's total number of hospital visits by 1.
  - *Shift function for updating opioid prescriptions:* The amount of opioid prescription the patient received is sampled from a normal distribution whose mean is equal to the average amount of opioid the patient received every time historically. The type of opioid the patient received is uniformly selected based on the frequency in the history. The standard deviation of the normal distribution is determined by the parameter, *shift function mode–prescription*. Similarly, "Expectation" has a std of 0, "Normal" has a small std, and "Aggressive" has a large std.
  - *Projection function:* At every time step, we sample patients from the environment. For each patient, the probability of selection follows a Normal distribution with a mean of 1% and standard deviation of 0.2%. We simulate the feedback model for 500 steps.
- **Results:** The complete regression results for the opioids risk case study are shown in Table II. The exact configurations used for the tradeoff analysis in the radar plot of the paper are as follows:
  - Configuration b1: *ML model* = MLP, *doctor threshold* = 0.7, *shift function mode-hospital* = aggressive, *shift function mode-prescription* = aggressive.
  - Configuration b2: *ML model* = XGBoost, *doctor threshold* = 0.7, *shift function mode-hospital* = normal, *shift function mode-prescription* = expectation.
  - Configuration b3: *ML model* = MLP, *doctor threshold* = 0.5, *shift function mode-hospital* = aggressive, *shift function mode-prescription* = aggressive.

### *C. Predictive Policing Case Study*

Given historical crime incident data for a collection of regions, the task is to decide how to allocate patrol officers to the areas. We leveraged the predictive policing dataset from [10] to study the long-term fairness impact of the ML-enabled system.

- **ML model:** We used the SEPP model used by [10], which is the well-known crime prediction model used by PredPol software and the study conducted by Lum and Isaac [11]. We leveraged the crime dataset generator created by Akpinar et al., which is based on the crime incidents in Bogotá [10]. The generator produces the location and time

TABLE II: The complete coefficients and sum of squares of opioid risk scoring case study.

| Terms | Dummies | Coeffs | Sum Sq. | $\eta^2$(%) |
|---|---|---|---|---|
| (Intercept) | | 2.14E-02*** | 3.26E-05 | 0.49 |
| *ML model* | xgboost | 1.12E-02*** | 6.35E-03*** | 96.80 |
| *Doctor threshold* | | -6.10E-05 | 5.51E-06*** | 0.08 |
| *Shift function–hospital* | normal | 3.00E-04 | 3.92E-05*** | 0.60 |
| | aggressive | 5.70E-05 | | |
| | equal | -3.03E-04 | | |
| *Shift function–prescription* | normal | -2.37E-03*** | 3.82E-05*** | |
| | aggressive | -2.38E-03*** | | 0.59 |
| *(ML model, doctor threshold)* | xgboost | -4.74E-04*** | 9.44E-06*** | 0.14 |
| *(ML model, Shift function–hospital)* | (xgboost, normal) | -4.15E-04* | 1.92E-05*** | 0.29 |
| | (xgboost, aggressive) | -3.21E-04 | | |
| | (xgboost, equal) | -1.77E-03*** | | |
| *(ML model, Shift function–prescription)* | (xgboost, normal) | 2.53E-03*** | 6.28E-05*** | |
| | (xgboost, aggressive) | 2.65E-03*** | | 0.96 |
| *(Doctor threshold, Shift function–hospital)* | normal | 6.00E-05 | 1.54E-06 | 0.02 |
| | aggressive | 3.60E-05 | | |
| | equal | 2.47E-04* | | |
| *(Doctor threshold, Shift function–prescription)* | normal | 5.80E-05 | 9.60E-08 | |
| | aggressive | 3.50E-05 | | 0.00 |
| *(Shift function–hospital, Shift function–prescription)* | (normal, normal) | 1.10E-04 | 1.28E-06 | 0.02 |
| | (aggressive, normal) | 3.14E-04 | | |
| | (equal, normal) | 2.37E-04 | | |
| | (normal, aggressive) | 4.50E-05 | | |
| | (aggressive, aggressive) | -2.08E-04 | | |
| | (equal, aggressive) | 9.50E-05 | | |

***p $<$ .001; **p $<$ .01; *p $<$ .05

TABLE III: The complete coefficients and sum of squares of predictive policing case study.

| Terms | Coeffs | Sum Sq. | $\eta^2$(%) |
|---|---|---|---|
| *(Intercept)* | 3.06E-01*** | 4.78E-02 | 26.17 |
| *Discovery rate–hotspot* | 8.99E-03*** | 8.48E-03*** | 4.64 |
| *Discovery rate–other* | -3.46E-02*** | 1.26E-01*** | 68.92 |
| *hotspot effect area range* | -7.18E-04 | 5.40E-05 | 0.03 |
| *(Discovery rate–hotspot, discovery rate–other)* | 1.75E-03 | 3.21E-04 | 0.18 |
| *(Discovery rate–hotspot, hotspot effect area range)* | -5.36E-04 | 3.00E-05 | 0.02 |
| *(Discovery rate–other, hotspot effect area range)* | 9.40E-04 | 9.30E-05 | 0.05 |

***p $<$ .001; **p $<$ .01; *p $<$ .05

of 270,352 crime incidents spread over 19 districts for 2,190 days. For long-term analysis, we further used the data generation model from [10] to generate crime data for 2400 days, which includes 315,687 crime incidents in total. Following their pipeline, we used the data of 500-2000th days for the model training, and 2000-2400th days for the simulation. Each district is divided into cells and SEPP model computes the crime intensity of each cell for the next day based on past data.

- **Decision maker/policy:** Based on the crime intensity of each cell predicted by the SEPP model, following the policy of [10], we select the 50 cells with the highest crime intensities as hotspots. Extra police force will be allocated to the hotspots on the next day.
- **Environment:** The environment is represented as a dataset of incidents, storing when and where each incident will take place. Specifically, the 2000-2400th days' incidents of the generated data is used to represent the environment.
- **Dynamics:** The selected hotspot cells are patrolled by more police. Hence, the crime discovery rates in the hotspots are higher than in the other cells. We used a range of discovery rates, [0.8, 1.0] and [0.2, 0.5] in hotspot and non-hotspot cells, respectively. The discovery rate is the probability of every crime incident happening in that cell that can be discovered. Furthermore, [10] used $1km \times 1km$ as the patrolling area size of hotspot cells. However, the designers may want to explore the long-term fairness impact of this setting. Therefore, we used the hotspot effect area range as a configurable parameter as well, with the options [1, 3], representing $1km \times 1km$, $2km \times 2km$ and $3km \times 3km$, respectively. The shift function will simulate if the incidents of the current day will be discovered according to the police allocation. The projection function will save the discovered incident information in the police system for future prediction.
- **Results:** The complete regression results for the predictive policing case study are in Table III. The exact configurations used for the tradeoff analysis in the radar plot of the paper are as follows:
  - Configuration c1: *Discovery rate–hotspot* = 1.0, *Discovery rate–other* = 0.5, *hotspot effect area range* = 3.
  - Configuration c2: *Discovery rate–hotspot* = 0.95, *Discovery rate–other* = 0.5, *hotspot effect area range* = 3.
  - Configuration c3: *Discovery rate–hotspot* = 0.80, *Discovery rate–other* = 0.5, *hotspot effect area range* = 3.

# References

[1] A. D'Amour, H. Srinivasan, J. Atwood, P. Baljekar, D. Sculley, and Y. Halpern, "Fairness is not static: deeper understanding of long term fairness via simulation studies," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 2020.

[2] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," in *Advances in neural information processing systems*, 2016.

[3] L. T. Liu, S. Dean, E. Rolf, M. Simchowitz, and M. Hardt, "Delayed impact of fair machine learning," in *ICML*, 2018.

[4] R. Parthvi, A. Agrawal, S. Khanijo, A. Tsegaye, and A. Talwar, "Acute opiate overdose: an update on management strategies in emergency department and critical care unit," *American journal of therapeutics*, 2019.

[5] E. A. Mensah, M. J. Rahmathullah, P. Kumar, R. Sadeghian, and S. Aram, "A proactive approach to combating the opioid crisis using machine learning techniques," in *Advances in Computer Vision and Computational Biology*. Springer, 2021, pp. 385–398.

[6] J. S. Hastings, M. Howison, and S. E. Inman, "Predicting high-risk opioid prescriptions before they are given," *Proceedings of the National Academy of Sciences*, vol. 117, no. 4, pp. 1917–1923, 2020.

[7] R. Vunikili, B. S. Glicksberg, K. W. Johnson, J. T. Dudley, L. Subramanian, and K. Shameer, "Predictive modelling of susceptibility to substance abuse, mortality and drug-drug interactions in opioid patients," *Frontiers in Artificial Intelligence*, vol. 4, p. 742723, 2021.

[8] A. E. Johnson, L. Bulgarelli, L. Shen, A. Gayles, A. Shammout, S. Horng, T. J. Pollard, S. Hao, B. Moody, B. Gow *et al.*, "Mimic-iv, a freely accessible electronic health record dataset," *Scientific data*, 2023.

[9] G. A. Adam, C.-H. K. Chang, B. Haibe-Kains, and A. Goldenberg, "Hidden risks of machine learning applied to healthcare: unintended feedback loops between models and future data causing model degradation," in *Machine Learning for Healthcare Conference*, 2020.

[10] N.-J. Akpinar, M. De-Arteaga, and A. Chouldechova, "The effect of differential victim crime reporting on predictive policing systems," in *Proceedings of the ACM FAccT*, 2021.

[11] K. Lum and W. Isaac, "To predict and serve?" *Significance*, 2016.