Introduction to Resubmission

C1: Scope seemed too vast for a Phase 1 project, which can dilute the likelihood of success for achieving all three. These topics are coupled at some level, but achieving all three aims may be too optimistic for a Phase 1 effort. Proposing too many innovations would dilute the likelihood of success in achieving all three.

Our response: the scope of this proposal has been revised to focus mainly on Fairness (bias in healthcare-related prediction models) in liver transplants; however, we still propose to provide an explainable AI module in Phase 2. Additionally, the new stature guidelines for the current Funding Opportunity Announcement Number (PA-22-176) allow us to double the Award Project Period to 1 year from an initial six months proposed in our original submission, allowing our team to propose a less time-constrained project timeline.

C2: The proposal has no specific timelines and a breakdown of the tasks to complete each specific aim.

Our response: It will take our team six months to complete each Aim. We have provided specific timelines for this resubmission and outlined a milestone breakdown of the tasks to complete each specific aim in Phase 1. In addition, we have included a description of how the Phase 1 work fit into the larger vision of Phase 2.

C3: There is mention of potential organ/donor characteristics that could be important, but no discussion is provided in the significance section. Regarding the Research Strategy component of the proposal, it was very surprising that the "Significance" section was only a single paragraph, which is quite a small portion of the Research Strategy allocated for this section.

Our response: the significance section has been expanded to include more comprehensive background information. We have included a discussion of a preliminary analysis of liver donation across 330 hospitals in the US. Additionally, we include more details on how organs are currently assigned and emphasize the instances where the MELD score doesn't currently correlate to a higher chance of receiving a liver. Finally, we introduce **FairMatch**, a data-driven alternative to the MELD score with the potential to achieve high liver transplant success while preserving fairness.

C4: There are mentions of public datasets and data collection from hospitals for Aim 1, but the only dataset they have access to is from STAR (total size, plans for training, and testing split). Where will the other data sets come from? A brief discussion would benefit the proposal.

Our response: we have included a brief discussion on the characteristics of the Standard Transplant Analysis and Research (STAR) dataset, our chosen dataset for our proof-of-concept work in Phase 1. We want to stress that the STAR dataset contains information from 330 hospitals across the US and a sample of 286,082 patients' health records.

C5: It is unclear what other co-PIs contribute to the program's implementation. Most of the implementation seems to be dependent on the PI.

Our response: our small business has secured a collaboration with co-investigators Xiaoqian Jiang and Nathan R. Hoot from UTHealth. Their background in biomedical informatics and liver transplantation is key to the success of our project. Together, we will co-design a fairness-first prediction engine behind FairMatch to help surgeons determine the best liver transplant candidate for each donor's liver. Their contribution to the project will be divided into two phases. During our Phase 1 work, we will build a fairness-first engine to predict donor and recipient compatibility based on each patient's medical history data. In Phase 2, we will work with our UTHealth partners (co-investigators) to develop a human-AI interface to visualize the model interpretation while collecting feedback from medical doctors to improve FairMatch.

C6: Although model-fairness can be difficult to achieve, as investigators noted to balance out the fairness and prediction accuracy. The solution that investigators suggested does not suffice for the problems they identified. Furthermore, it should consider that the marginalized populations may not have been fairly represented in the dataset they plan to use.

Our response: In this resubmission, we elaborate on our proposed approach based on an embedding layer to handle sparse data and a distillation network containing distilled information from a tree-based model to handle dense features, which will improve its accuracy in Phase 2 of this proposal. At the same time, the system we present in this resubmission is a pioneer in providing fairness in liver transplantation with a data-driven approach. While Phase 1 will allow us to link fairness and prediction accuracy, Phase 2 will help us improve the system as new expert feedback and labeled data become available. Finally, we want to highlight that the STAR dataset is the most suitable resource for liver transplant information for our proposed work with a reasonable patient representation of 286,082 medical records.

C7: Unclear if the solution accounts for and differentiates between MELD 3.0. A brief discussion on how it differs from this newly published approach would further strengthen the innovation.

Our response: The MELD 3.0 score is still mainly based on statistical analysis and heuristic rules as its predecessors. One of its key limitations is that it is difficult to adapt to new data with different distributions. Though new MELD score versions have explicitly considered gender bias, it is still hard to extend to other kinds of bias, e.g., race, age, etc. More detail on FairMatch'sdistillation and tree-based model approach is provided in this resubmission.

Specific Aims

End-stage liver disease and fulminant hepatic failure can be treated by a liver transplant. However, transplanted organ supplies are limited compared to the number of patients on the waiting list [1]. As a result, organ assignment becomes a critical decision that requires significant thought, particularly when considering the ethical principles of utility (maximization of net benefit to the community) and justice (the fair distribution pattern of benefits). The Model for End-stage Liver Disease (MELD) score evaluates the patient's current state based on three lab test findings, including serum creatinine, serum sodium, total bilirubin, and INR of prothrombin time, is a widely used assignment technique [2]. A higher MELD score indicates that a patient's status is worsening and that the patient has a greater priority for organ transplantation. MELD 3.0 gives a more accurate prediction of mortality in general than MELDNa and addresses the determinants of waitlist outcomes, including sex disparity [3].

Despite its popularity, the MELD score has two major disadvantages. First, the MELD score does not clearly consider the post-transplant outcome [2], which is an essential indicator for organ distribution decisions. The MELD score shows a very poor link with graft survival rate across genders and races. Second, the MELD score excludes organ and donor characteristics [2], which may lead to suboptimal organ allocation. As a result, researchers are encouraged to develop new substitution assignment systems for liver transplants [4].

To better forecast post-transplant outcomes, machine learning (ML) has developed data-driven solutions for the organ transplant challenge. The main concept is to train an ML model that takes patient and donor characteristics as input and predicts outcomes such as pre-transplant mortality, post-transplant mortality, and so on. The trained algorithm is then used to predict a score for each patient-donor pair, which can assist physicians in making organ donation choices. Several ML models have recently been implemented and show potential in the organ transplant problem [5, 6], such as employing logistic regression and gradient boosting models to predict mortality in liver transplant recipients [7], using neural networks and random forests to predict graft failure after transplant [8], or interpretable systems for real-time organ allocation [9].

Unfortunately, current research suggests that ML models in organ transplantation may be unjust and biased towards specific categories of people. Earlier research has addressed such questions of fairness [10, 11]. However, whereas fairness issues in machine learning have lately received a lot of attention [12], there have been few attempts to research the fairness issue in organ transplant jobs. Due to two impediments, developing a fair ML system with competitive accuracy for organ transplants remains difficult. To begin, organ transplant datasets comprise dense and sparse categorical information (e.g., numerical lab test results) (e.g., blood type of recipients and donors). Existing research merely employs one-hot encoding to transform sparse characteristics [13]. However, owing to the curse of dimensionality, one-hot encoding may result in unacceptable performance when the feature cardinality is high. Second, including fairness goals in the training process is difficult. Previous research has mostly used tree-based models [14] for organ transplant prediction because of their superior performance in dealing with dense inputs. On the other hand, existing bias mitigation methods are primarily concerned with the training process, including loss design and representation learning [15], neither of which can be directly applied to tree-based models due to the indifferentiable trait.

AI POW LLC is a Texas-based company that integrates automation and interpretability technologies for ML. To address these issues, the AI POW team proposes **FairMatch**: An Automated Patient-Organ Assigning System Tailored to Fairly Predicting Failure Rate in Liver Transplant. The AI POW team has partnered with McGovern Medical School at UTHealth in a joint interest in proposing a robust prediction framework for liver transplant graft failure, one of the most critical post-transplant outcomes. AI POW utilizes an embedding layer to handle sparse data and a distillation network containing distilled information from a tree-based model to handle dense features, motivated by DeepGBM's [16] high performance in recommendation tasks. This architecture combines the benefits of tree-based models with deep neural networks in handling sparse and dense information and allows us to use in-processing debiasing techniques to ensure fairness. The AI POW-UTHealth partnership will develop a debiasing technique that addresses fairness difficulties in both the knowledge distillation and end-to-end training stages. Extensive studies on the Standard Transplant Analysis and Research (STAR) [17] dataset will illustrate the superiority of our system.

The AI POW LLC team and the McGovern Medical School at UTHealth seek to achieve the following:

Aim 1. Months 0-6: Collect and pre-process organ transplant datasets, and develop machine learning models to accurately predict the failure rates based on patient and organ features in a data-driven manner.

Aim 2 Months 6-12: Mitigate the model unfairness with regularization techniques to ensure that patients in different sensitive groups (e.g., race and gender) are treated fairly.

The goals of Phase I are to develop a method for integrating the accuracy of predictive analytics with a fairness-centric approach. In this initial proof-of-concept, the AI POW team will build on the preliminary results of previous systems that have shown accurate results. In Phase II, our proposed transplant system will offer an explainability module to aid model credibility among clinicians and expand its reach nationally.

Research Strategy

(A) Significance

A liver transplant is the only curative treatment for end-stage liver disease. A liver transplant can treat various conditions, ranging from hepatitis B and C to primary biliary cirrhosis, autoimmune hepatitis, Wilson disease, hemochromatosis, and inborn errors of metabolism. For patients with end-stage liver failure due to chronic viral hepatitis B or C infection, who are not candidates for antiviral therapy, liver transplantation may reduce or eliminate fatigue and other symptoms associated with their disease.

Whether or not to proceed with a liver transplant depends on many factors, including the availability of organs and patient characteristics and ability. A growing medical challenge is that demand for donor livers far exceeds supply. For this, the matching process between donor organs and patients is complex and time-consuming, particularly because patient-organ matching involves multiple criteria, including medical urgency, size compatibility, blood type compatibility, and geographic proximity to the donor organ center. The United Network for Organ Sharing (UNOS) is a non-profit organization that operates the organ transplant system in the US. UNOS manages the organ allocation system to ensure that organs are distributed fairly among patients who need them. However, recent studies have shown that this existing automated matching system has several limitations. In particular, system outcomes may be biased against certain categories of people.

The Model for End-stage Liver Disease (MELD) score measures liver function, and it's a prevalent method used to prioritize patients for a liver transplant in the US. The MELD score is calculated using three lab values: serum creatinine (mg/dl), the international normalized ratio of prothrombin time, and bilirubin. Some versions of this score, the MELDNa, include serum sodium for calculation [18], while the newer MELD 3.0 provide more accurate mortality measures than its predecessors. For pediatric patients, a separate score called Pediatric End-stage Liver Disease (PELD) score is used [19]. Based on the definition of the MELD score, patients with higher MELD scores should have a higher priority in matching the organs.

The limitations of this scoring system are well-documented. It does not consider important post-transplant outcomes such as patient survival or early death after transplantation, nor does it account for the age or sex of either patients or donors, which can influence outcomes in ways that elude our ability to measure them accurately at this time. And while these scores are based on doctors' expertise, such a heuristic approach is prone to human bias, not to mention that it is difficult to adapt such scores for new distributions of patient and donor data. The complexity of organ allocation has led many experts to call for a more scientific approach that maximizes the net benefit to society while supporting a fair distribution pattern of benefits. The limitations of current methods in predicting graft failure in liver transplants motivate the development of new machine learning (ML) models. Unfortunately, these new models may be unfair to some groups of people. **A fair organ allocation system is important because it ensures that organs are distributed based on medical needs, not other factors.**

The fairness problem in liver transplants can be defined following established practices in the medical field. We examine fairness at the group level concerning race and gender. A fair liver distribution system should give the patients on the waiting list under the same medical condition the same chance of receiving a suitable organ [20, 21, 22]. **However, fairness is subjective, so equal chances could have different interpretations. We consider fairness from two perspectives in this proposal.** Consideration 1, it is expected that patients in different groups will have an equal percentage of being predicted as graft failed. Because of this, if organs are allocated based on predicted scores, patients in different groups will tend to receive them equally. Consideration 2, ML models are expected to provide an equal prediction quality for different groups, which can be quantified by true positive and false positive rates of graft failure prediction.

The aforementioned fairness definitions correspond to two commonly used fairness metrics for ML models: demographic parity and equalized odds. Demographic parity requires that different groups have an equal rate of positive outcomes, and equalized odds require that the true positive rate and false positive rate be equal. We quantify the degree to which these metrics are met using demographic parity difference (DPD) and equalized odds difference (EOD), respectively—a method that builds on previous work [23].

Based on the definition of MELD score, patients with higher MELD scores should have a higher priority in receiving an organ, which contradicts our findings. The Standard Transplant Analysis and Research (STAR) dataset consolidates data from 330 hospitals across the US and 286,082 patients registered on the Organ Procurement and Transplantation Network (OPTN) waiting list. In Figure 1, using the STAR dataset, our team provides a population size and average MELD score across races and genders, while in Figure 2, we provide an average organ receiving rate and graft failure rate across races and genders. From Figure 1 (left), we can determine that while Race V has the highest reported Average MELD score of 27.5, this group presented one of the lowest Organ Received Rates (.48) in Figure 2

(left). In contrast, Race VII with an average MELD score of 20.4, reported the highest Organ Received Rate (.62) in our analysis. Such biases are also observed among genders; while Female patients show a higher Average MELD score in Figure 1 (right), **such a score doesn't correlate to a higher chance of receiving a liver as compared to their male counterpart group**, as displayed in Figure 2 (right).

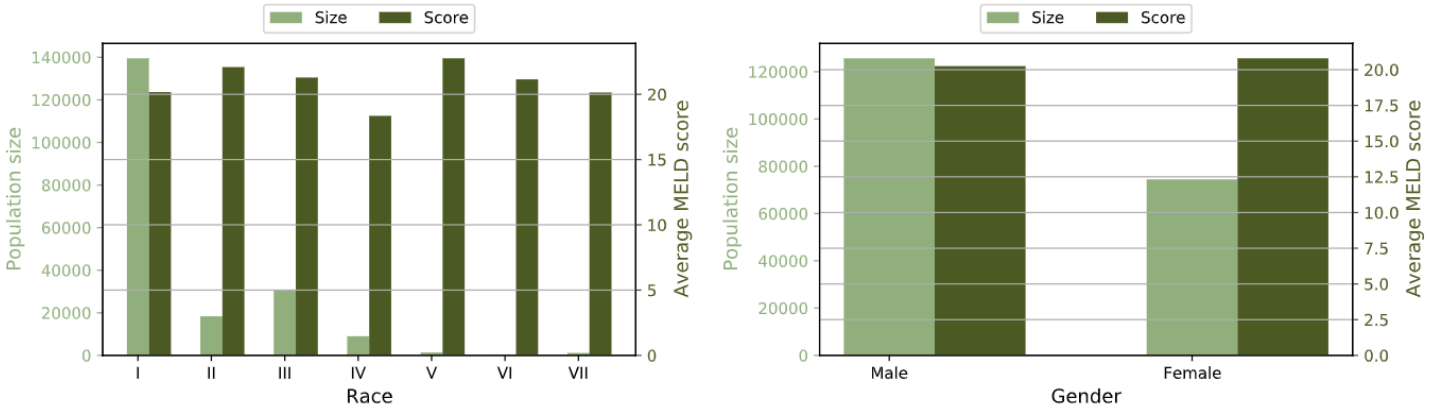

Figure 1. Population size and average MELD score across races and genders.

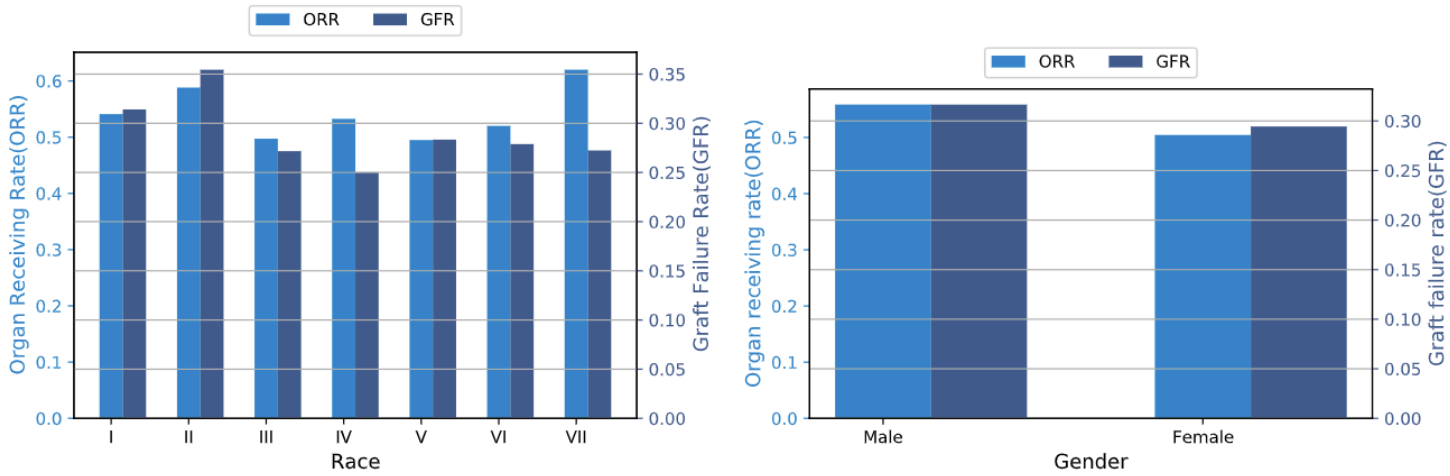

Figure 2. Average organ receiving rate and graft failure rate across races and genders.

Additionally, despite its widespread use, the MELD score has several drawbacks. First, it is still inflexible to consider comprehensive post-transplant mortality metrics such as Graft Failure Rate—an important factor for organ distribution decisions. Second, it ignores patient, donor, and organ characteristics; thus, it may lead to inappropriate allocation decisions. Consequently, the MELD score may be an inefficient tool for clinicians when prioritizing their patients on the waiting list for a liver transplant. For this, we propose a new automated organ-patient matching system, **FairMatch**, a data-driven solution with the potential to achieve high liver transplant success while preserving fairness. Our team envisions that by the end of Phase 2, FairMatch will provide medical doctors with an ML-based system for a fair and efficient organ allocation without requiring them to have a background in computer science.

(B) Innovation: Machine learning (ML) has provided data-driven solutions for the organ transplant task to better model post-transplant outcomes. At the core of FairMatch, there is a technology that trains an ML model that takes the features of patients and donors as input and outputs the predicted failure rate. Then, the trained model is deployed to predict a score for each patient-donor pair, intended to help clinicians make decisions about organ transplants. To train a strong machine learning model, our team proposes an embedding layer to handle the sparse features and a distillation network with distilled knowledge from a tree-based model to handle the dense features. This design can combine the advantages of tree-based models and deep neural networks in handling sparse and dense features.

Our team will develop bias mitigation techniques based on AI POW's prediction technologies to enforce the model to make fair predictions. Despite the strong performance of ML models, they could be unfair and show bias against certain groups of people, such as races and genders, leading to a negative social impact on the decisions. To tackle the unfairness issues, AI POW's technology uses fairness regularization losses to force the model to make fair predictions for different groups of patients. AI POW further devises a two-step debiasing strategy that mitigates the fairness issues in

both the knowledge distillation and the end-to-end training stages. In summary, AI POW's proposed technology will enable a comprehensive machine learning solution for accurate failure rate prediction that considers fairness.

	MELD-score
Organ receiving rate	-0.32376
Graft failure rate	0.36653

Table 1. Correlation between MELD score and Organ receiving and Graft failure rate.

(C) Approach

In 2021 CEO Alfredo Costilla presented this concept to the Liu idea lab for innovation and entrepreneurship in Houston, Texas, where he was accepted into Rice University's Innovation accelerator to further FairMatch's value proposition. In February 2022, Dr. Costilla's team raised a pre-seed round of capital from private investors and proposed an early prototype with collaborators at UTHealth. As part of Dr. Alfredo's Rice Innovation fellowship, he has been awarded a $15,000.00 grant for the preliminary work the project proposed here. Additionally, Dr. Alfredo has secured a formal collaboration with industry experts at UTHealth to effectively form world-class cross-disciplinary work. Particularly, Co-investigators Xiaoqian Jiang (biomedical informatician) and Nathan R. Hoot (a physician who provides care for liver transplantation) will help in this project to further develop, test, and improve the prediction and explainability engine behind the liver Patient-Organ Matching feature described in this proposal. As part of this, the AI POW LLC team is implementing a lean methodology to further elaborate on the development of this work. To achieve this, the team validates or invalidates business and technical hypotheses with medical doctors facing challenges in liver transplantation procedures.

While the team has enough funds to cover the Phase I to Phase II time gap, the AI POW team is planning to raise a seed round of fundraising in 2023, which will be critical for this project, such capital will be needed as the team plans to expand the team to further develop a front-end and back-end accessible to third parties of the project presented here. Thus, the team expects to have in place matching funds for the development of this project. In addition to the team's previous experience securing private funding sources, we also count on the experience of preparing, securing, executing, and successfully delivering results for SBIR grants (NSF SBIR Award 2136679), which put us in the best position to execute the proposed work here.

As part of our preliminary studies on the MELD scores, we compare them with machine learning approaches to validate the proposed system. Following our analysis of Figures 1 and 2, we calculate the Pearson correlation between organ receiving rate and MELD, presented in Table 1. A score of **-0.32376** means that the MELD score has no close relation with the organ receiving rate from the group-level analysis. Similarly, a score of **0.36653** implies the MELD score cannot indicate group-level graft failure rate at the post-transplant stage. In Table 2, AI POW conducts graft failure rate prediction with ML approaches and MELD score. AI POW can find that the performance of ML approaches in terms of ROC AUC (higher is better) is significantly higher than the MELD score and is also capable of identifying fairness issues in the data (through computing two fairness metrics: DPD and EOD [24]).

Model	Sensitive attribute: Race			Sensitive attribute: Gender		
	ROC AUC	DPD	EOD	ROC AUC	DPD	EOD
MELD-score	0.505±0.000	—	—	0.505±0.000	—	—
Logistic Regression	0.777±0.000	0.648±0.017	0.834±0.007	0.777±0.000	0.021±0.000	0.033±0.001
Random forest	0.804±0.000	0.630±0.030	0.703±0.047	0.804±0.000	0.020±0.001	0.036±0.001
GBDT	0.809±0.000	0.637±0.027	0.713±0.033	0.809±0.000	0.017±0.000	0.031±0.001

Table 2: Comparison between MELD-score and ML approaches.

Phase I Work Plan: The Co-founder and CEO of AI POW LLC will serve as a PI; he and his project team will focus on further validating the feasibility of the predictive ML framework for liver transplant and the explainability features and fitness. In addition, the project team includes co-investigator Xiaoqian and co-investigator Nathan to bring cross-disciplinary expertise from the medical domain. Finally, expert Kwei-Herng 'Henry' Lai will lead our efforts in predicting liver transplant graft failure, one of the most important post-transplant outcomes. AI POW plans to validate the proposed framework through extensive experiments on the STAR dataset. This organ transplant dataset is collected from 286,082 patients registered on the OPTN waiting list and de-identified by removing all the identifiers from the data and randomly shifting dates under IRB protocol approval (HSC-MS-13-0549). The STAR dataset is the most suitable resource for liver transplant information for our proposed work with a reasonable patient representation of 286,082 medical records that contain the biomedical information of organ recipients and organs/donors. The study's subjects include those on a

waiting list for organ transplants and recipients who received organs in the US between April 1, 2000, and February 28th, 2013, across 330 hospitals. The dataset also records the outcomes of recipients' post-transplant treatments through December 31, 2013. Of a subset of 160,360 recipients we analyzed, 41.8% experienced graft failure within a six-month period following transplantation. We plan to select 40 features from recipients and 40 features from organs/donors. The race and gender of each recipient will be marked as confidential information.

Aim 1: Collect and pre-process organ transplant datasets, and develop machine learning models to accurately predict the failure rates based on patient and organ features in a data-driven manner. In this task, we aim to develop machine learning models to predict patient outcomes, such as pre-transplant and post-transplant mortality, in a data-driven manner. Figure 3 shows an overview of the workflow. Firstly, we will introduce how we will collect and pre-process the raw data. Then, our team elaborates on handling the sparse and dense features with tree-based models and deep neural networks.

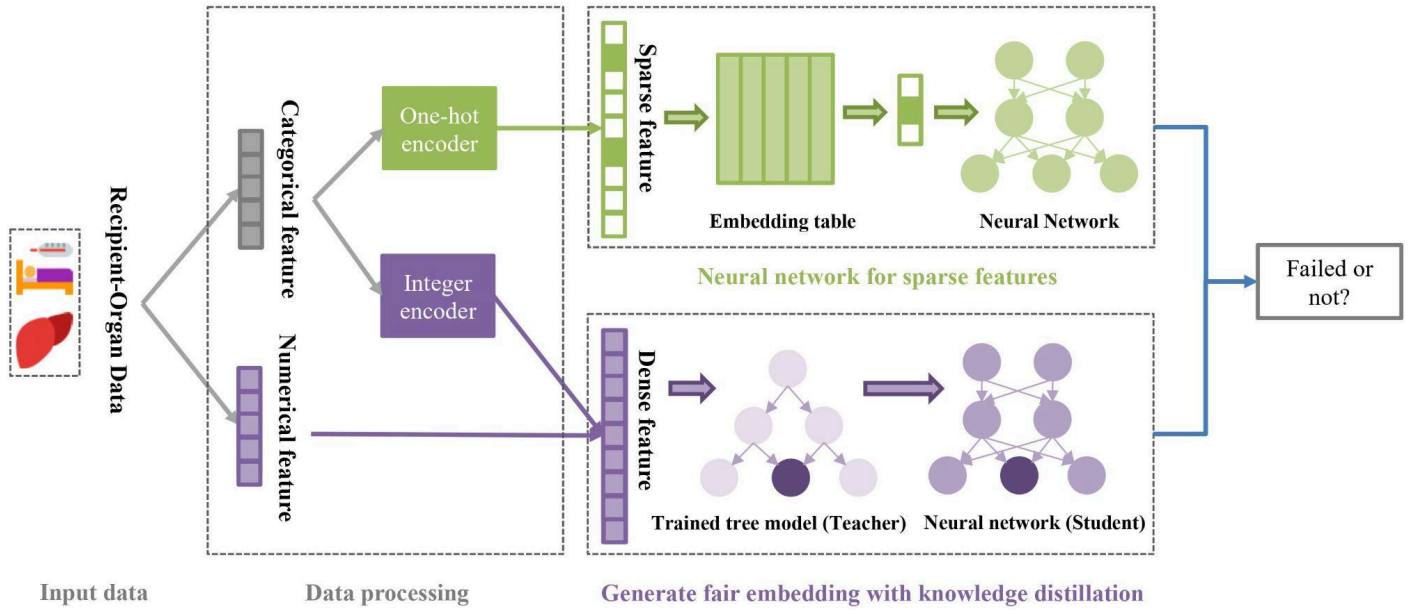

Figure 3. An overview of the patient outcome prediction framework.

Data pre-processing (2-month task). AI POW will leverage both public datasets and collect data from the hospitals to construct the datasets. As mentioned earlier, AI POW plans to use the STAR as our public organ transplant dataset. AI POW will also connect with hospitals to construct private datasets, following a similar procedure as above. Following the data pre-processing practice in machine learning, AI POW first imputes the missing values. Specifically, AI POW uses zeros to replace the missing values for the numeric data. Then, AI POW identifies the categorical features (i.e., the features that only have a fixed number of values) and numerical features from the recipient and organ features. For the categorical features, AI POW employs two kinds of encoders, including a one-hot encoder that maps the raw features to one-hot sparse vectors and an integer encoder that transforms the categorical features into numerical values, where the latter are further concatenated with the original numerical features to serve as the final dense features.

Combining deep learning and tree-based models for patient outcome prediction (4-month task). Motivated by the strong performances of the tree-based methods on dense features and deep learning models on handling sparse features via embedding tables, AI POW proposes to combine them to achieve the best prediction results. Here is how our team plans to handle the following two kinds of features. **(1) Sparse features:** The sparse features from the recipient and the organ are combined and processed by a categorical neural network, an embedding lookup layer that maps categorical indices to dense vectors, followed by feature interactions. Then a factorization machine (FM) is adopted to learn the first/second-order interactions of these features, and a deep neural network is applied to learn the higher-order interactions of these features. The output of FM and the neural network are summed to obtain the final sparse representations. **(2) Dense features:** our team plans to combine the dense features of the recipient and organ. Our team proposes to train a neural network to distill the knowledge from a trained tree-based model to take advantage of the tree-based models in handling dense features. This is not an easy task because the structures of the trees and neural networks are naturally different. Motivated by [16], AI POW proposes an effective tree distillation strategy by distilling the clustering patterns of the leaf nodes. First, since tree-based methods often do not use all the features but instead greedily

choose the useful features, AI POW will only select the used features of a tree to train the neural network. Second, AI POW will train a neural network by distilling the knowledge of how the tree partitions the data. Specifically, a tree-based model essentially partitions the data into different clusters, where the data in the same leaf node belong to the same cluster. AI POW trains the weights of dense networks to distill the knowledge from such tree structures. **Final representations:** The final output is obtained by combining sparse and dense representations, followed by several fully connected layers to make predictions. AI POW will leverage cross-entropy loss to train the network in a highly modular and end-to-end fashion.

Aim 1: Milestone, Potential Pitfalls, and Alternative Strategies. The milestone for Aim 1 will focus on the development of the patient outcome prediction model and the tuning of the model design. The potential pitfall is that the model could overfit the data as the datasets could not be large enough, which means the model may perform well on the training data but suffer from a performance decrease on the testing data. AI POW will first explore different regularization terms to alleviate the overfitting issue if this happens. Additionally, AI POW will try trimming down the model size so that the model will not overfit the training data. AI POW will also try to collect more data so that our team can train more robust models on the datasets. Another potential pitfall is that it could be difficult to select the proper features since not all the features are useful. If this happens, a potential strategy is to use feature selection algorithms to automatically filter out the most important features for model training.

Aim 2: Mitigate the model unfairness with regularization techniques to ensure that patients in different sensitive groups (e.g., race and gender) are treated fairly. In this task, AI POW aims to mitigate the unfairness in the model. AI POW will first introduce the potential fairness issues in the model. Then AI POW presents fairness losses to enforce the model to make fair predictions. Finally, AI POW proposes a tailored two-step debiasing strategy to achieve model fairness via debiasing the distillation network and the prediction network.

Fairness problem. While machine learning models can make accurate precisions, they can be biased and show discrimination against certain groups of people [10,11,24]. AI POW approach focuses on fairness at the group level and race and gender groups. Specifically, the predictor should allow patients of different races and genders to have an equal chance of receiving compatible organs. However, fairness is subjective, so equal chances could have different interpretations. In this proposal, AI POW considers fairness defined from two perspectives. On the one hand, AI POW expects the patients in different groups to have an equal percentage of being predicted as “failed.” In this sense, patients in different groups will tend to equally receive an organ if allocating organs based on the predicted score. On the other hand, machine learning models are expected to provide an equal prediction quality for different groups, which can be quantified by true positive and false-positive rates of the prediction.

Setting fairness metrics in FairMatch (2-month task). The above two fairness definitions correspond to two commonly used fairness metrics for machine learning models: demographic parity and equalized odds, where the former demands different groups to have an equal percentage of a positive outcome, and the latter requires equal true positive and false-positive rates. Specifically, AI POW will set the mechanisms in the prediction framework to quantify the degrees of demographic parity and equalizes odds with demographic parity difference (DPD) and equalized odds difference (EOD) [24]. DPD is obtained by calculating the maximum differences in the ratios of positive outcomes across groups. EOD can be similarly obtained by the maximum differences between true positive and false-positive rates.

Fairness loss implementation (2-month task). AI POW proposes to use fairness loss to enforce the model to make fair predictions. The key idea is to enforce all the sensitive attributes to have similar prediction distributions to the majority group. Specifically, the loss is obtained by subtracting the expected predictions of all the patient-organ pairs from the expected predictions of the majority group. In this way, the predictions of all the groups will be forced to be similar to those of the majority group. AI POW will apply the mini-batch to train the deep learning models.

Two-step debiasing (2-month task). To maximally debias the model, AI POW proposes a two-step debiasing strategy to debias both the categorical neural network and the network for dense features. In the first step, AI POW achieves fair knowledge distillation by plugging in the fairness loss into the distillation objective. In the second step, AI POW incorporates the fairness constraint into the end-to-end training. These two debiasing steps complement each other towards final fair predictions. The first step focuses on the dense representations that serve as the input of the end-to-end training, and the second step debiases the CatNN and the embedding tables.

Aim 2 Milestone, Potential Pitfalls, and Alternative Strategies. The milestone for Aim 2 will focus on the implementation of the fairness losses for the distillation training and the end-to-end training. A potential pitfall is that achieving fairness and prediction accuracy could be challenging, and they have a tradeoff. In this situation, AI POW will use a parameter to control this balance so that customers can deploy the model based on their fairness needs. Another potential pitfall is that other fairness goals could be beyond demographic parity and equalized odds. If this happens, AI POW will design new fairness losses based on the fairness needs.

Phase II - Look Ahead.

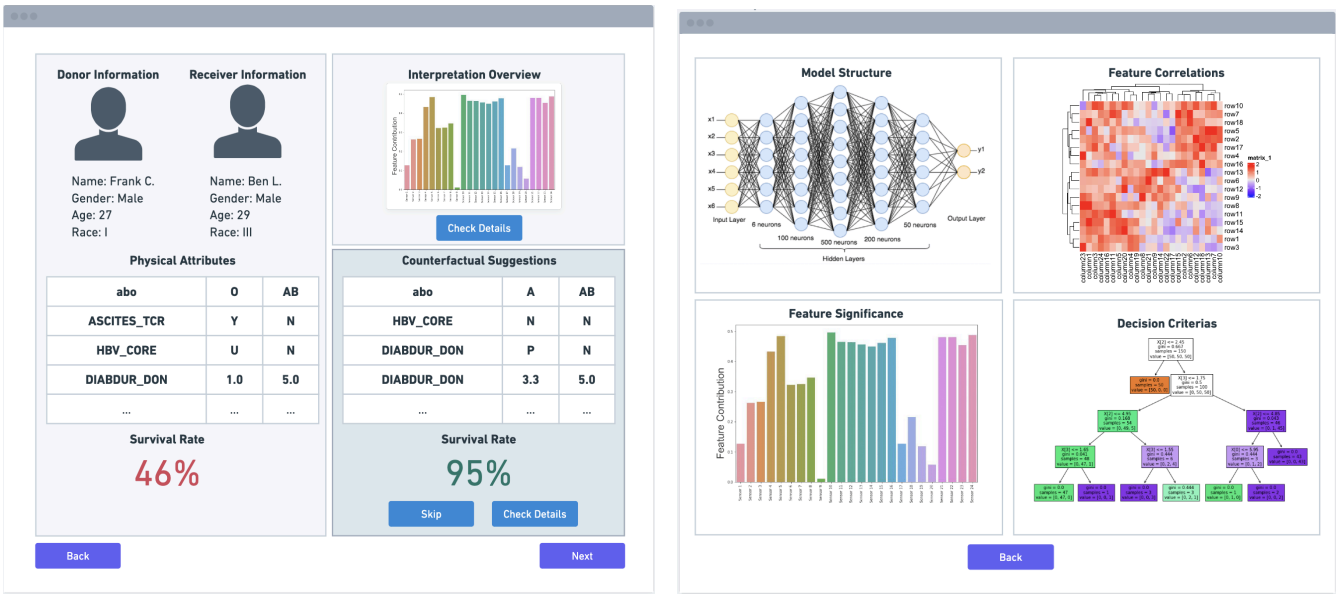

Figure 4: Human-AI Interface (Left: Prediction Summary; Right: Model Interpretation)

Human-AI Interface.

Since visual representation is more intuitive for humans to better understand complex information, in Phase 2 of this proposal AI POW will develop a human-AI interface to visualize the model interpretation while collecting feedback from medical doctors to enhance the model explainer. Specifically, we consider that the human-AI interface will have two main pages: prediction summary and model interpretations. Figure 4 shows the wireframe design of how our team envisions the human-AI interface. **Prediction Summary:** On this page, AI POW aims to provide an outcome prediction summary by presenting the donor and receiver's physical attributes and the predicted graft failure rate. In addition, AI POW's explainable AI module will provide an interpretation overview to visualize the significance of key features that lead to this prediction. The counterfactual suggestion section provides potential insights into the surgical procedure by providing counterfactual instances. In **FairMatch 2.0,** physicians will have a "skip" button to see the next counterfactual instance and a "check details" button to elaborate on the current instance. By clicking "skip," **FairMatch 2.0** will be able to collect feedback from medical doctors on invalid instances and further improve the underlying interpretation algorithm.

Additionally, clicking the "Check Details" button in the counterfactual suggestion panel will lead to the prediction summary page that compares and provides model interpretations on the counterfactual instance. **Model Interpretation:** By clicking the "Check Details" in the panel of "Interpretation Overview," the model interpretation will show up to provide details, including model structure, feature correlations, feature significance as well as decision criteria of the model. Users can also click on each panel to get further details.

Summary. At the end of Phase 1, AI POW will have implemented the complete workflow of the failure rate prediction model and developed a prototype of fairness modules. AI POW will also have received feedback from their users, i.e., doctors in the hospital, and evaluated their effectiveness. Our team plans to validate the proposed framework through extensive experiments on the STAR dataset. Phase 2 efforts will focus on extending the scope of the FairMatch technology by developing an interpretability module with our vision for a Human-AI interface. Additionally, Phase 2 will require developing human-in-the-loop technology to interactively improve the prediction, model interpretation, and a larger implementation of the system at the center and national level.

Commercialization. The global market for organ transplantation is expected to become $61.5 billion by 2027, expanding at a growth rate of 9.2% [25] in the same period, driven by increasing incidences of organ failures and rising demand for transplant products such as tissue products, immunosuppressants, and organ preservation solutions. Only in 2021 in the US were performed 9,236 liver transplants [26] with an average cost of US$163,438 [27]. Such cost puts tremendous pressure on the patient, their families, and the US healthcare system. Therefore, an efficient and fair liver organ-patient matching is a critical issue that costs the US healthcare system millions of dollars annually. A tool to better allow doctors to make educated decisions during such important surgeries will save lives and reduce costs associated with liver transplants. This project is an ambitious approach to fairness-ML in a clinical setting; for this, the AI POW team has partnered with UTHealth to co-develop the FairMatch framework proposed for phase 1. As a small business, this partnership will be key in producing a positive impact for more hospitals, physicians, patients, and their families.

References:

[1] Abouna, G.M., 2008, January. Organ shortage crisis: problems and possible solutions. In Transplantation proceedings (Vol. 40, No. 1, pp. 34-38). Elsevier.

[2] Wiesner, R., Edwards, E., Freeman, R., Harper, A., Kim, R., Kamath, P., Kremers, W., Lake, J., Howard, T., Merion, R.M., and Wolfe, R.A., 2003. Model for end-stage liver disease (MELD) and allocation of donor livers. Gastroenterology, 124(1), pp.91-96.

[3] Kim, W. R., Mannalithara, A., Heimbach, J. K., Kamath, P. S., Asrani, S. K., Biggins, S. W., Wood, N. L., Gentry, S. E., & Kwong, A. J. (2021). MELD 3.0: The Model for End-Stage Liver Disease Updated for the Modern Era. Gastroenterology, 161(6), 1887–1895.e4. https://doi.org/10.1053/j.gastro.2021.08.050

[4] Merion, R.M., Wolfe, R.A., Dykstra, D.M., Leichtman, A.B., Gillespie, B. and Held, P.J., 2003. Longitudinal assessment of mortality risk among candidates for liver transplantation. Liver transplantation, 9(1), pp.12-18.

[5] Yoon, J., Alaa, A., Cadeiras, M. and Van Der Schaar, M., 2017, February. Personalized donor-recipient matching for organ transplantation. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 31, No. 1).

[6] Berrevoets, J., Jordon, J., Bica, I. and van der Schaar, M., 2020. OrganITE: Optimal transplant donor organ offering using an individual treatment effect. Advances in neural information processing systems, 33, pp.20037-20050.

[7] Byrd, J., Balakrishnan, S., Jiang, X. and Lipton, Z.C., 2021. Predicting mortality in liver transplant candidates. In Explainable AI in Healthcare and Medicine (pp. 321-333). Springer, Cham.

[8] Lau, L., Kankanige, Y., Rubinstein, B., Jones, R., Christophi, C., Muralidharan, V. and Bailey, J., 2017. Machine-learning algorithms predict graft failure after liver transplantation. Transplantation, 101(4), p.e125.

[9] Berrevoets, J., Alaa, A., Qian, Z., Jordon, J., Gimson, A.E. and Van Der Schaar, M., 2021, July. Learning Queueing Policies for Organ Transplantation Allocation using Interpretable Counterfactual Survival Analysis. In International Conference on Machine Learning (pp. 792-802). PMLR.

[10] Bertsimas, D., Papalexopoulos, T., Trichakis, N., Wang, Y., Hirose, R. and Vagefi, P.A., 2020. Balancing efficiency and fairness in liver transplant access: tradeoff curves for the assessment of organ distribution policies. Transplantation, 104(5), pp.981-987.

[11] Parent, B. and Caplan, A.L., 2017. Fair is fair: We must re-allocate livers for transplant. BMC medical ethics, 18(1), pp.1-7.

[12] Du, M., Yang, F., Zou, N. and Hu, X., 2020. Fairness in deep learning: A computational perspective. IEEE Intelligent Systems, 36(4), pp.25-34.

[13] Bishara, A.M., Lituiev, D.S., Adelmann, D., Kothari, R.P., Malinoski, D.J., Nudel, J.D., Sally, M.B., Hirose, R., Hadley, D.D. and Niemann, C.U., 2021. Machine Learning Prediction of Liver Allograft Utilization From Deceased Organ Donors Using the National Donor Management Goals Registry. Transplantation direct, 7(10).

[14] Indyk, P. and Motwani, R., 1998, May. Approximate nearest neighbors: towards removing the curse of dimensionality. In Proceedings of the thirtieth annual ACM symposium on Theory of computing (pp. 604-613).

[15] Caton, S. and Haas, C., 2020. Fairness in machine learning: A survey. arXiv preprint arXiv:2010.04053.

[16] Ke, G., Xu, Z., Zhang, J., Bian, J. and Liu, T.Y., 2019, July. DeepGBM: A deep learning framework distilled by GBDT for online prediction tasks. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (pp. 384-394).

[17] Procurement, O., & Network, T. (2017). Standard Transplant Analysis and Research (STAR) Dataset Files.

[18] Biggins, S. W., Kim, W. R., Terrault, N. A., Saab, S., Balan, V., Schiano, T., Benson, J., Therneau, T., Kremers, W., Wiesner, R., Kamath, P., & Klintmalm, G. (2006). Evidence-based incorporation of serum sodium concentration into MELD. Gastroenterology, 130(6), 1652–1660. https://doi.org/10.1053/j.gastro.2006.02.010

[19] McDiarmid, S. V., Merion, R. M., Dykstra, D. M., & Harper, A. M. (2004). Selection of pediatric candidates under the PELD system. Liver transplantation: official publication of the American Association for the Study of Liver Diseases and the International Liver Transplantation Society, 10(10 Suppl 2), S23–S30. https://doi.org/10.1002/lt.20272

[20] Good, M. D., James, C., Good, B. J., & Becker, A. E. (2005). The culture of medicine and racial, ethnic, and class disparities in healthcare. The Blackwell companion to social inequalities, 396-423.

[21] Egede, L. E. (2006). Race, ethnicity, culture, and disparities in health care. Journal of general internal medicine, 21(6), 667.

[22] Hamberg, K. (2008). Gender bias in medicine. Women's health, 4(3), 237-243.

[23] Bird, S., Dudík, M., Edgar, R., Horn, B., Lutz, R., Milan, V., Sameki, M., Wallach, H. and Walker, K., 2020. Fairlearn: A toolkit for assessing and improving fairness in AI. Microsoft, Tech. Rep. MSR-TR-2020-32.

[24] Kaufman, S.R., 2013. Fairness and the tyranny of potential in kidney transplantation. Current Anthropology, 54(S7), pp.S56-S66.

[25] iHealthcareAnalyst, Inc., November 8, 2021 (Accessed April 3, 2022).
https://www.ihealthcareanalyst.com/global-organ-transplantation-market/#:~:text=The%20global%20market%20for%20organ,immunosuppressants%2C%20and%20organ%20preservation%20solutions.
[26] United Network for Organ Sharing (UNOS), Jan 28, 2022 (Accessed April 3, 2022).
https://unos.org/news/in-focus/2021-9th-record-year-liver-transplants/
[27] Van der Hilst, C.S., IJtsma, A.J., Slooff, M.J. and TenVergert, E.M., 2009. Cost of liver transplantation: a systematic review and meta-analysis comparing the United States with other OECD countries. Medical Care Research and Review, 66(1), pp.3-22.

Narrative

End-stage liver disease and acute liver failure, such as hepatic failure, can be treated by a liver transplant. However, compared to the number of patients on the waiting list, transplanted organ supplies are limited. Only in 2021, the US performed 9,236 liver transplants with an average cost of US$163,438, representing tremendous pressure on the patient, their families, and the US healthcare system. AI POW LLC proposes FairMatch, a fair and automated patient-organ assigning system tailored to predict failure rates in liver transplants.

Equipment

No special equipment is required to complete the Aims outlined in this Phase I proposal. Items required to complete the Aims are outlined in Materials and Supplies in the Budget Justification.

Abstract

End-stage liver disease and fulminant hepatic failure can be treated by a liver transplant. However, compared to the number of patients on the waiting list, transplanted organ supplies are limited. As a result, organ assignment becomes a critical decision that requires significant thought. AI POW LLC is a Texas-based company that integrates automation and interpretability technologies for ML. To address these issues, the AI POW team proposes FairMatch, An Automated Patient-Organ Assigning System Tailored to Fairly Predicting Failure Rate in Liver Transplant. The AI POW team has partnered with McGovern Medical School at UTHealth in a joint interest in proposing a robust prediction framework for liver transplant graft failure, which is one of the most critical post-transplant outcomes. AI POW utilizes an embedding layer to handle sparse data and a distillation network containing distilled information from a tree-based model to handle dense features, motivated by DeepGBM's high performance in recommendation tasks. This architecture combines the benefits of tree-based models with deep neural networks in handling sparse and dense information and allows us to use in-processing debiasing techniques to ensure fairness. The AI POW-UTHealth partnership will develop a debiasing technique that addresses fairness difficulties in both the knowledge distillation and end-to-end training stages. Extensive studies on the Standard Transplant Analysis and Research (STAR) dataset will illustrate the superiority of our system. This project seeks to achieve the following **Aims: 1.** Collect and pre-process organ transplant datasets, and develop machine learning models to accurately predict the failure rates based on patient and organ features in a data-driven manner. In this task, we aim to develop machine learning models to predict patient outcomes, such as pre-transplant and post-transplant mortality, in a data-driven manner. Firstly, we will introduce how we will collect and pre-process the raw data. Then, AI POW elaborates on how to handle the sparse and dense features with tree-based models and deep neural networks. **2**. Mitigate the model unfairness with regularization techniques to ensure that patients in different sensitive groups (e.g., race and gender) are treated fairly. In this task, AI POW aims to mitigate the unfairness in the model. AI POW will first introduce the potential fairness issues in the model. Then AI POW presents fairness losses to enforce the model to make fair predictions. Finally, AI POW proposes a tailored two-step debiasing strategy to achieve model fairness via debiasing the distillation network and the prediction network.

Inclusion of Women and Minorities

Women and minorities will be included in the study. We will make every effort to recruit and enroll women and minority participants at a level that is *at least* equivalent to the proportion they are represented in our eligible population.

Inclusion of Individuals Across the Lifespan

Patients. All patients, as long as they are 18 years of age or older, are eligible for this study.

PROTECTION OF HUMAN SUBJECTS

Source of data

We will use simulated and de-identified data to conduct our proposed research. We expect this project to be eligible for NIH Exempt Human Subjects Research under §CFR 46.104(d)(4) because the collection/study of data is recorded such that subjects cannot be identified. Our research teams are highly educated, trained, and experienced in legal, regulatory, and ethical requirements and best research practices involving human data. We will maintain robust confidentiality protections, including standard technical safeguards to remove identifiers from our study.

Interaction with patients

The proposed study is focused on methodology development and will use only existing genomics databases. Any form of interaction with the patients or intervention will not take place. Based on these characteristics of the study, our self-assessment is that the study presents no more than minimal risk to human subjects. Nevertheless, we will seek IRB guidance on whether the study is classified as having a minimal risk and whether a waiver of consent would be warranted given the impracticality of seeking consent from the targeted patient group.

Potential risks

The data sets that are used for method development are from existing data. The primary potential risk to subjects is loss of confidentiality with potential linkage to external data to reveal identity. Given our team's previous experience of handling such sensitive data and a strong background in health privacy, this is unlikely to happen. Our project will not reveal any patient-level information.

Adequacy of Protection Against Risk

This study will use retrospectively collected data to study fairness-enhancing methods for AI models in the healthcare context. Although the investigation has little risk, we consider the privacy risk as being yet unknown, so we will still maintain a high standard of protection in our study

- The servers will be located behind firewalls in our institutions.
- Access to the development servers will be limited to the approved study investigators and their collaborators.

Data and Safety Monitoring Plan

In addition to the security layers deployed by the IT Security team at UTHealth, the School of Biomedical Informatics has a strong and dedicated IT team that maintains the privacy and security of relevant research. All SBMI servers are hosted in a private HIPAA-compliant environment to provide privacy-preserving storage and high-performance computing capabilities to biomedical and behavioral researchers. The data collected and generated in this project will be stored and processed in this secure SBMI computing environment.

The research team will review any issues with data safety in a regular meeting and modify or stop the research protocol if necessary. We will abide by all the rules and plans developed for human subject protection that apply to this proposed project.

Potential benefits

Our study has the potential to enhance data sharing and inform better decision-making for transplantation. These results are valuable in promoting and generating new scientific evidence for biomedical research. The discoveries from this project may not directly benefit individuals whose data will be used for the analyses. Overall, the risks to subjects are reasonable concerning the potential benefit to research participants and others.

We will abide by all the rules and plans developed for human subject protection that apply to this proposed project.