Algorithm, Human, or the Centaur: How to Enhance Clinical Care?

Arlen Dean
Saïd Business School, Oxford University, Oxford, OX11HP, UK, arlen.dean@sbs.ox.ac.uk

Agni Orfanoudaki
Saïd Business School, Oxford University, Oxford, OX11HP, UK, agni.orfanoudaki@sbs.ox.ac.uk

Soroush Saghafian
Harvard Kennedy School, Harvard University, Cambridge, MA 02138, USA, soroush_saghafian@hks.harvard.edu

Harini A. Chakkera
Division of Transplantation, Mayo Clinic Hospital, Phoenix, Arizona, 85054, USA, Chakkera.Harini@mayo.edu

Curtiss B. Cook
Division of Endocrinology, Mayo Clinic Arizona, Scottsdale, Arizona, 85259, USA, Cook.Curtiss@mayo.edu

Problem definition: Machine learning (ML) algorithms are increasingly used to enhance clinical care for a wide range of medical conditions. However, clinical decision-making often involves nuanced, contextual factors that ML models cannot easily account for, suggesting unrealized potential in combining human expertise with ML. This raises critical questions about how human expertise can complement ML models and the extent to which such collaboration could enhance performance in practice. **Methodology/results:** We propose a human-algorithm "centaur" model to harness synergies between human experts and ML. Through rigorous theoretical analysis, we motivate our model's design and characterize its advantages over standalone ML and human-only approaches in predictive tasks. We introduce a framework for operationalizing centaur models in practice and validate our approach using a case study with a leading U.S. hospital focused on predicting 30-day readmissions in transplantation patients. Our findings consistently demonstrate that the centaur model can outperform both human experts and the best ML algorithm. **Managerial Implications:** We show that even when the accuracy of human prediction is low, creating a centaur model can yield substantial performance improvements by systematically enhancing ML algorithms with human insight. Furthermore, the predictive accuracy of the centaur model improves proportionally with the degree of divergence between the human-only and ML-only estimation policy, highlighting the value of capturing complementary information from both sources. Finally, our empirical results reveal that (1) while ML models excel in identifying non-linear and dynamic risk patterns, human experts tend to rely on linear assessments driven by primarily static, time-invariant features; (2) on average, human experts tend to over-estimate risks compared to ML models; and (3) centaurs can potentially impact clinical practices by addressing a major barrier in implementing pure ML algorithms: their lack of alignment with human experts' clinical intuition, and hence, providing advice that can receive high weights by practitioners.

Key words: Machine Learning, Human-Algorithm Interactions, Healthcare, Transplantation, Readmission, Hospital Operations

1. Introduction

Machine learning (ML) algorithms are poised to provide practitioners and hospital administrators with novel data-driven tools that could improve patient prognosis, clinical diagnosis, and treatment as well as hospital efficiency (Beam and Kohane 2018). Yet, there is an "inconvenient truth" about machine learning (ML) in healthcare (Panch et al. 2019). The nuances of human judgment needed in patient care are too complex for any data-driven model to capture completely (Kelly et al. 2019). Given healthcare's high-stakes and human-centric nature, practitioners remain cautious about relying solely on algorithms for critical decisions (Luan et al. 2024). Consequently, while ML models are increasingly employed to generate recommendations, final decisions are entrusted to trained clinical practitioners, underscoring the indispensable role of human oversight in patient-centered care.

The critical awareness of ML's limitations extends beyond just the institutional level. Physicians often exhibit reluctance to trust algorithmic recommendations in critical decisions, even when evidence shows these recommendations are more accurate than their own. This skepticism arises particularly when algorithmic outputs conflict with their medical intuition (Jussupow et al. 2020), a phenomenon commonly referred to as "algorithm aversion" (Dietvorst et al. 2015). As a result, the effectiveness of ML-based decision support tools in clinical practice remains, to a large extent, a function of the behavioral characteristics of its users and their willingness to accept algorithmic suggestions (Dai and Singh 2021). Overcoming these challenges requires rethinking how ML models can effectively integrate human intuition and expert insights. This leads to a critical question: *How can ML systems be designed to effectively learn from and complement the cognitive and intuitive frameworks of human experts?*

In this paper, we propose incorporating the clinical reasoning of human experts into the algorithm training and validation process. Unlike standard human-in-the-loop approaches, our method is characterized by two important features: symbiotic learning and direct incorporation of human intuition into algorithmic models (Saghafian and Idan 2024). We, therefore, term this hybrid approach the "centaur" model, drawing inspiration from the unique style of playing chess known as centaur chess, where humans and machines collaborated as a single team (Goldstein et al. 2017, Saghafian and Idan 2024). As chess grandmaster Garry Kasparov observed: "Weak human plus machine plus better process was superior to a strong computer alone and, more remarkably, superior to a strong human plus machine plus inferior process." (Kasparov 2010). This statement motivates the potential of utilizing the centaur model for care delivery. Specifically, by enabling ML systems to incorporate nuanced, context-specific insights from clinicians, we hypothesized that centaur models can address algorithm aversion, improve trust, and enhance predictive accuracy in high-stakes scenarios. These hybrid models represent a promising paradigm for advancing patient-centered, data-driven decision-making in healthcare.

1.1. Paper Overview

We formalize our proposed *centaur* model by offering a theoretical framework that enables systematically integrating human expertise within ML algorithms to enhance predictive accuracy in clinical decision-making and improve uptake by practitioners. Through rigorous theoretical analysis, we characterize the performance of the centaur model, highlighting its benefits over (a) standalone human experts, (b) pure machine learning models, and (c) alternative approaches that incorporate human input into ML algorithms. Building on this theoretical groundwork, we propose a generalizable framework to operationalize centaur models in clinical practice. We validate the framework through an extensive case study in the context of solid-organ transplantation, focusing on the critical task of predicting 30-day readmission. Collaborating with medical experts at the Mayo Clinic Arizona, we utilize a comprehensive dataset encompassing 1,537 kidney, liver, and heart transplantations and conduct a prospective survey to derive and evaluate our models.

The transplantation case study allows us to explore the interplay between human intuition and ML precision. First, we investigate the relative predictive accuracy of human experts versus ML algorithms in the task of 30-day readmission. Second, we evaluate whether a centaur model can outperform the best standalone ML algorithm, demonstrating the potential benefits of our proposed class of models. Third, we analyze the reasoning patterns of human experts and ML algorithms, assessing (a) whether they prioritize similar clinical features, and (b) if clinical experts systematically overestimate risk compared to algorithms. Fourth, we explore how an ML tool influences human experts' risk perception in the presence of algorithm aversion. Finally, we quantify the economic value of implementing a centaur-enhanced risk assessment process in clinical practice, estimating the cost savings associated with more accurate identification of at-risk patients.

1.2. Paper Contributions

Our study offers the following contributions:

- We formalize the *centaur* model as a hybrid framework that integrates human intuition with ML algorithms. Our analysis demonstrates that a centaur model systematically outperforms naive methods of combining human and ML predictions, such as simple averaging or ensemble stacking. Unlike these methods, which rely on fixed or pre-determined weights, the centaur model dynamically incorporates human predictions as a feature, allowing the model to adapt the influence of human input based on its value relative to the task. This approach achieves lower mean squared error by leveraging the *complementarity* between human experts' intuition and ML algorithms predictions.
- We demonstrate that the centaur model yields clear benefits when human predictions provide information not encoded in the ML model's features or predictions, such as tacit knowledge

unique to human expertise. It also achieves superior performance when ML predictions are highly accurate but incomplete, as the human input often contributes orthogonal or supplementary information that enhances the model's predictive capacity. However, the centaur's advantage diminishes when human predictions are either strongly correlated with the ML predictions or very poorly aligned with the true outcome, as the model gains limited additional value from redundant or misaligned inputs.

- We propose a generalizable framework to operationalize centaur models in practice and validate it through a real-world case study. Using retrospective data from a leading U.S. hospital, we develop, to the best of our knowledge, the first ML model capable of accurately predicting 30-day readmission across any major solid organ (kidney, liver, and heart). Additionally, we collect prospective survey data to capture medical experts' predictions, allowing us to compare their performance with the ML model. Our analysis shows that while the ML model outperforms human experts in terms of predictive accuracy, the two approaches rely on different clinical features, highlighting their complementary strengths. We then follow our centaur approach and find that integrating human intuition with the ML algorithm systematically improves predictive performance. Thus, we validate our theoretical findings with empirical evidence, showing that centaur models can outperform both the algorithm and the human experts, even though the human experts' performance is lower on average than the pure ML model. Leveraging these results, we show how making use of the centaur-enhanced risk assessment process could lead to reduced operational costs in clinical practices.
- Finally, we find that human experts tend to (a) overestimate risks compared to ML models, and (b) they rely mainly on linear relationships between some main features, whereas ML models effectively capture nonlinear relationships across a broader range of variables. Put together, our findings indicate that human and ML approaches bring fundamentally different strengths to risk assessment, and combining these through a centaur model offers a powerful way to enhance predictive accuracy, reduce costs, and support more effective decision-making in practice.

The remainder of the paper is organized as follows. Section 2 provides a summary of the literature relevant to our study. Section 3 introduces our centaur model and presents our main theoretical insights. Section 4 outlines a framework for developing, implementing, and validating human-algorithm centaur models in practice. In Section 5, we present our case study setup and results, and in Section 6, we summarize the managerial insights gained from our work. Finally, we conclude in Section 7 with an overview of the key findings and limitations of our work.

2. Literature Review

Three main streams of literature are particularly relevant to our study: (1) human-in-the-loop approaches that aim to augment algorithm recommendations with human guidance; (2) human-AI interaction studies on how humans benefit from algorithm suggestions; (3) medical studies on 30-day readmission after solid-organ transplantation.

Human-in-the-loop. Human-in-the-loop methodologies aim to integrate human feedback into the deployment of ML models. These approaches often assume that ML starts off less accurate than human experts (Wu et al. 2022). However, in healthcare, the opposite is frequently true: state-of-the-art ML models currently outperform human experts on average (McKinney et al. 2020, Tian et al. 2024). Furthermore, the integration of human feedback is often hindered by the practical challenges posed by existing healthcare infrastructure, which limits the feasibility of interactive or active ML approaches (Holzinger 2016, Panch et al. 2019). This raises the question of whether and how to best leverage human input when ML already achieves superior accuracy (Reardon 2019).

Alternate approaches, such as the two-way personalization model proposed by Saghafian (2024), integrate clinician preferences into causal inference algorithms, enabling the personalization of treatment recommendations for both patients and physicians. Similarly, Ibrahim et al. (2021) introduced a system to elicit human judgment for prediction algorithms, assuming that experts have at their disposal subject information unavailable in the model input. Arnold et al. (2019) conducted a prospective observational study providing evidence that combining human and ML model insights could enhance discrimination performance. Building on these ideas, Choudhary et al. (2023) and Peng et al. (2024) explored human-AI ensembles, offering managerial insights and analytical evidence on their potential benefits. In contrast to the aforementioned works, we propose a new framework to integrate expert intuition into ML systems in the form of an exogenous predictive model that is trained on both human judgment and typical training data sets. Our proposed centaur framework, thus, differs from the existing approaches, creating a more effective way of combining algorithms with human expert intuition (see also Saghafian and Idan (2024)).

Human-AI Interactions. Our work directly complements the literature on navigating human-AI collaboration. Various metrics to quantify the weight of advice related to human judgment in the context of algorithms have been proposed (Harvey and Fischer 1997, Bailey et al. 2022). Grand-Clément and Pauphilet (2024) derived theoretical insights into the impact of adherence in algorithmic guidance. To deepen the understanding of how algorithmic advice is adopted, McLaughlin and Spiess (2022) investigate the role of reference points in shaping human decision-making, while Athey et al. (2020) focus on the cost of effort associated with following algorithmic recommendations. Caro and de Tejada Cuenca (2023) and Sun et al. (2022) use historical data on worker compliance to refine algorithm designs, aiming to reduce deviations from data-driven recommendations.

Logg et al. (2019) provide empirical evidence showing that people can prefer algorithmic recommendations to human judgment. Boyacı et al. (2024) show that machine-based recommendations can improve the overall accuracy of human decisions, although their effect is less pronounced when humans are under cognitive load. Balakrishnan et al. (2022) explore the role of private information, uncovering that a weighted average model can capture how humans incorporate advice into their predictions. Imai et al. (2020) developed a general-purpose statistical method that can experimentally evaluate the causal impact of algorithmic suggestions on human decisions. Kawaguchi (2021) show through an empirical analysis that humans are more likely to follow algorithmic recommendations when their forecasts are integrated into the algorithm. In the clinical domain, several articles have examined how liability and task uncertainty shape the adoption of algorithmic recommendations (Dai and Singh 2021, Bertsimas and Orfanoudaki 2021). Our study extends this literature by examining how integrating human input can systematically improve ML accuracy and exploring this dynamic through theoretical modeling and empirical validation in a healthcare setting.

Extending the existing literature, our model builds on the concept of stacking (Breiman 1996, LeBlanc and Tibshirani 1996), treating human and ML outputs (e.g., readmission risk) as separate features for weighted predictions. However, unlike traditional stacking, our approach embeds the human expert's output directly within the ML model as a new feature alongside other variables (e.g., patient features). As a result, our model requires a different set of techniques to evaluate its potential benefits in downstream performance. In this regard, our analysis aligns with feature selection studies (Chen 1976, Muthukrishnan and Rohini 2016, Beraha et al. 2019), which typically analyze model performance by excluding features based on predefined criteria (e.g., mutual information score). Unlike these works, the criterion in our setting is determined by the ML model itself, which adjusts the weight assigned to the embedded feature.

Transplantations. Our study complements the medical literature focusing on early readmissions (defined as occurring within 30 days after discharge) after solid organ transplantation. Readmissions, often linked to index admission factors (Patel et al. 2016), are costly and risky for transplantation patients and hospitals, making their reduction a key priority and quality measure for hospitals and health systems (Jencks et al. 2009). Improving quality measures has grown more critical as public reporting of medical outcomes gains traction in efforts to enhance transparency in care (Saghafian and Hopp 2020). Lubetzky et al. (2016), Orfanoudaki et al. (2023), and King et al. (2017) study readmissions following kidney transplantation. Patel et al. (2016), Munshi et al. (2020), and Werner et al. (2016) focus on liver transplantation, while Bachmann et al. (2018), Jalowiec et al. (2008) and Goyal et al. (2021) study readmission factors for heart transplantation. Common patient features, such as metabolic factors and blood glucose (BG) management, have been highlighted as important variables during the immediate period after a transplant for both

kidney and liver patients (Boloori et al. 2015, Chakkera et al. 2009, Munshi et al. 2020). To the best of our knowledge, there has not been any study that directly explores the commonalities of these factors and others across solid-organ transplantations in the context of early hospital readmissions. Our work proposes, for the first time, a multi-organ early readmissions risk prediction method after transplantation, introducing one coherent model for kidney, liver, and heart transplant patients.

3. The Centaur Model

In this section, we formally introduce the human-algorithm centaur model and rigorously characterize its potential advantages in practical applications. To this end, we assume that human expert risk assessments are generated independently, without knowledge of the ML model predictions. The centaur model then integrates these human estimates to augment and refine its predictive ability.

3.1. Problem Notation

Given a data set of N total observations, we consider a vector of features $\mathbf{x}_i \subseteq \mathbb{R}^d$ and an outcome variable $y_i \subseteq \mathbb{R}$ for each patient $i \in [N]$. We let $f(\mathbf{x}_i) : \mathbf{x}_i \to \mathbb{R}$ represent an ML model that estimates the outcome y via an output denoted by $\hat{y}$. We also consider a model capturing the human experts' prediction function denoted by $l(\mathbf{x}_i) : \mathbf{x}_i \to \mathbb{R}$. To generate the hybrid centaur model, we incorporate $l(\mathbf{x}_i)$ as a new feature that augments the data set features $\mathbf{x}_i$. We then define our centaur as a new model $q(\mathbf{x}_i, l(\mathbf{x}_i)) : \{\mathbf{x}_i, l(\mathbf{x}_i)\} \to \mathbb{R}$. In reference to all observations, we use $\mathbf{X} \in \mathbb{R}^{N \times d}$ and $\mathbf{y} \in \mathbb{R}^{N \times 1}$. Table EC.1 summarizes the general notation of the manuscript.

In the following sub-sections, we theoretically study the performance of the proposed centaur model, using the expected mean-square error (MSE) as a main performance metric. We start by considering a stylized model that explicitly incorporates a learned representation of human assessments $l(\mathbf{X})$ as a feature. We then extend our analysis to highlight the regimes in which incorporating human knowledge in this way provides clear benefits over relying solely on the ML model $f(\mathbf{X})$. As we will show, such regimes include cases in which the performance of the pure ML model $f(\mathbf{X})$ is much stronger than that of the human experts' assessments $l(\mathbf{X})$. Thus, we establish that there is value in combining human intuition with an ML model, even if human intuition is weaker than the algorithm. As we will demonstrate, this is due to the centaur model's ability to effectively leverage the *complementarity* between the two information sources.

3.2. Motivating Modeling Choice

We motivate our choice of using the human expert's prediction as a feature by analyzing a baseline ML model of the following form:

$$f(\mathbf{x}_i) = \sum_{j=1}^{d} \sum_{m=1}^{M_j} \beta_{jm} \phi_m(\mathbf{x}_{ij}), \tag{1}$$

where $\phi_m(\mathbf{x}_{ij})$ is the mth basis function for observation i's jth feature, M_j is the number of basis functions corresponding to the jth feature, and coefficients β_{jm} are the parameters of the ML model. We next let the centaur be defined as

$$q(\mathbf{x}_i, l(\mathbf{x}_i)) = \theta_{d+1,1} l(\mathbf{x}_i) + \sum_{j=1}^{d} \sum_{m=1}^{M_j} \theta_{jm} \phi_m(\mathbf{x}_{ij}), \tag{2}$$

where θ_{jm} is the centaur model's parameter weights. We do not make any strict assumptions surrounding the functional form of the human's model $l(\mathbf{X})$ but assume, for simplicity, that it is a feature directly used without any transformation by the centaur. For ease of notation, we will reference $\theta_{d+1,1}$ as θ_{d+1}.

Using models (1) and (2), we first compare our approach to an ensemble learning approach, which involves averaging the human and ML predictions. That is, we consider a benchmark model

$$v(\mathbf{X}) = \alpha_1 f(\mathbf{X}) + \alpha_2 l(\mathbf{X}),$$

where α_1, α_2 are the ensemble model weights. The formulation of $v(\mathbf{X})$ in Proposition 1 follows the *ensemble* approach of stacking (Breiman 1996). In the literature, stacking is the most prominent way to incorporate predictions from multiple model sources. From a statistical learning perspective, ensemble learning is aimed at combining a mix of so-called weak models to produce a single stronger model. Our centaur model deviates from this framework by directly considering the human model as a feature supplementing the original observation features.

We first characterize the exact advantage of our centaur model over the naive ensemble case where $\alpha_1 = \alpha_2 = \frac{1}{2}$. Subsequently, we will extend our result to the more general case where weights α_1 and α_2 are general optimized values in [0,1] summing to 1.

PROPOSITION 1 **(Averaging versus Centaur)**. *Given ensemble model $v(\mathbf{X})$ with parameter weights $\alpha_1 = \alpha_2 = \frac{1}{2}$, it holds that*

$$\begin{aligned}
&\min_{\boldsymbol{\theta}} \left\{ E\left[\frac{1}{N} \sum_{n=1}^{N} \Big(y_i - q(\mathbf{x}_i, l(\mathbf{x}_i)) \Big)^2 \right] \right\} \\
&= \mathsf{MSE}(\mathbf{y}, v(\mathbf{X})) - \frac{1}{N} \sum_{n=1}^{N} \left(\mathbb{E}\Big[(f(\mathbf{x_i}) - l(\mathbf{x_i}))(\mathbf{y_i} - f(\mathbf{x_i})) \Big] - \left(\frac{f(\mathbf{x}_i) - l(\mathbf{x}_i))}{2} \right)^2 \right) - (\sigma_{l(\mathbf{x})} \theta^*_{d+1})^2,
\end{aligned}$$

where $\sigma^2_{l(\mathbf{X})}$ is the variance of $l(\mathbf{X})$ and θ^*_{d+1} is the least-square solution for $q(\mathbf{X}, l(\mathbf{X}))$. The term $\boldsymbol{\theta} \in \mathbb{R}^{d+1}$ denotes a vector representing the model parameters to be optimized for the centaur model and $\mathsf{MSE}(\mathbf{y}, v(\mathbf{X}))$ is the mean-square error between the outcomes $\mathbf{y}$ and the ensemble model's predictions $v(\mathbf{X})$).

The proof of this proposition and all other results can be found in Section EC.2 of the electronic companion. Proposition 1 demonstrates that the performance gap between the centaur and naive averaging critically depends on three factors: (1) the divergence between human and ML predictions, (2) the strength of the correlation between human predictions and the true outcome, and (3) the variance of $l(\mathbf{x})$. Specifically, our result characterizes the gap between the MSE corresponding to the least-square estimated $q(\mathbf{x}_i, l(\mathbf{x}_i))$ and that of the simple ensemble model (averaging human and ML model predictions). Our result highlights that this gap depends on the extent to which $f(\mathbf{x}_i)$ differs from $l(\mathbf{x_i})$ and y_i, as well as the strength of the relationship between $l(\mathbf{x_i})$ and y_i, as quantified by θ^*_{d+1}.

Proposition 1 does not immediately clarify whether the centaur model consistently outperforms ensemble averaging. To address this, we show in the following theorem that our method achieves a lower MSE for all convex combinations of the human expert's and ML predictions in the ensemble model $v(\mathbf{X}) = \alpha_1 f(\mathbf{X}) + \alpha_2 l(\mathbf{X})$, where $\alpha_1, \alpha_2 \geq 0$ and $\alpha_1 + \alpha_2 = 1$. This result establishes a bound on the stacking ensemble's performance, highlighting the centaur's flexibility in leveraging human knowledge compared to static averaging.

THEOREM 1 **(Stacking Ensemble MSE versus Centaur)**. *The MSE of the centaur model $q(\mathbf{X}, l(\mathbf{X}))$ is bounded by that of the stacking ensemble model $v(\mathbf{x}_i)$. Specifically, we have:*

$$\min_{\boldsymbol{\theta}} \left\{ E\left[\frac{1}{N} \sum_{n=1}^{N} \Big(y_i - q(\mathbf{x}_i, l(\mathbf{x}_i)) \Big)^2 \right] \right\} \leq \min_{\boldsymbol{\alpha}} \left\{ E\left[\frac{1}{N} \sum_{n=1}^{N} \Big(y_i - v(\mathbf{x}_i) \Big)^2 \right] \right\},$$

where the left-hand side is the optimized $\mathsf{MSE}\,(\mathbf{y}, q(\mathbf{X}, l(\mathbf{X})))$*, the right-hand side is the optimized* $\mathsf{MSE}\,(\mathbf{y}, v(\mathbf{X}))$*, and the term* $\boldsymbol{\alpha} = (\alpha_1, \alpha_2)$ *is the vector of parameter weights for the ensemble model.*

Theorem 1 highlights the centaur's advantage over stacking approaches by showing how incorporating human input directly as a feature enables the model to adapt dynamically to task-specific conditions. Unlike fixed-weight aggregation methods, the centaur's flexibility allows it to avoid the pitfalls of static weighting schemes, making it robust to variability in the quality and relevance of human input. As we show in the next section, this adaptability enables the centaur to better utilize complementary insights from human experts to enhance the ML model.

3.3. Characterizing Advantage of Human Input

We now proceed to characterize when having the human expert's knowledge $l(\mathbf{X})$ in our centaur model provides an added benefit over the use of a stand-alone ML model $f(\mathbf{X})$.

THEOREM 2 **(ML versus Centaur)**. *Let $\hat{l}(\mathbf{X})$ denote the ordinary least-square estimate of $l(\mathbf{X})$ using a model of the same functional form as $f(\mathbf{X})$. We have:*

$$\min_{\boldsymbol{\beta}} \left\{ E\left[\frac{1}{N} \sum_{n=1}^{N} \Big(y_i - f(\mathbf{x}_i) \Big)^2 \right] \right\} - \min_{\boldsymbol{\theta}} \left\{ E\left[\frac{1}{N} \sum_{n=1}^{N} \Big(y_i - q(\mathbf{x}_i, l(\mathbf{x}_i)) \Big)^2 \right] \right\}$$

$$= \left(\frac{1}{N} \sum_{i=1}^{N} (y_i - f(\mathbf{x}_i)) \Big(l(\mathbf{x}_i) - \hat{l}(\mathbf{x}_i) \Big) \frac{\sigma_{l(\mathbf{X})}}{\sigma^2_{M_{\mathbf{X}} \cdot l(\mathbf{X})}} \right)^2,$$

where the term $\sigma^2_{M_{\mathbf{X}} \cdot l(\mathbf{X})}$ is the variance of residuals for $M_{\mathbf{X}} \cdot l(\mathbf{X}) = l(\mathbf{X}) - \hat{l}(\mathbf{X})$, $\sigma_{l(\mathbf{X})}$ is the variance of $l(\mathbf{X})$, and $\boldsymbol{\beta} \in \mathbb{R}^d$ is the vector of parameters for the ML model.

Theorem 2 reveals that, besides the variance terms, the benefit of having $l(\mathbf{X})$ is highly dependent on both how well $f(\mathbf{X})$ predicts $\mathbf{y}$ (i.e., $y_i - f(\mathbf{x}_i)$) and how much information $l(\mathbf{X})$ provides that cannot be estimated via a model with $\hat{l}(\mathbf{X})$ (i.e., $l(\mathbf{x}_i) - \hat{l}(\mathbf{x}_i)$). While the result does not directly relate the centaur's performance to the accuracy of $l(\mathbf{X})$ in predicting y, we can use it to derive the following proposition that provides clarity on this question by characterizing some special cases.

PROPOSITION 2. *Let Δ denote the difference between minimal MSE of $f(\mathbf{X})$ and that of $q(\mathbf{X}, l(\mathbf{X}))$ as characterized in Theorem 2. We have:*

$$\Delta = \begin{cases} 0 & \text{if } f(\mathbf{x}_i) - l(\mathbf{x}_i) = 0,\ y_i - l(\mathbf{x}_i) \in \mathbb{R},\ \forall i \in [N] \\ \sigma^2_{l(\mathbf{X})} & \text{if } f(\mathbf{x}_i) - l(\mathbf{x}_i) \neq 0,\ y_i - l(\mathbf{x}_i) = 0,\ \forall i \in [N] \\ \left(\rho_{\mathbf{y}, l(\mathbf{X})} \cdot \mathsf{MSE}\Big(\mathbf{y}, f(\mathbf{X})\Big) \cdot \frac{\sigma_{l(\mathbf{X})}}{\sigma^2_{M_{\mathbf{X}} \cdot l(\mathbf{X})}} \right)^2 & \text{if } f(\mathbf{x}_i) - l(\mathbf{x}_i) \neq 0,\ y_i - l(\mathbf{x}_i) \neq 0,\ \forall i \in [N], \end{cases}$$

where $\rho_{\mathbf{y}, l(\mathbf{X})}$ is the correlation between $\mathbf{y}$ and $l(\mathbf{X})$.

The first case of Proposition 2 confirms that the centaur model's benefit diminishes as human predictions and ML estimates converge (i.e., $f(\mathbf{x}_i) - l(\mathbf{x}_i) \to 0$). This is true regardless of the accuracy of human predictions relative to the true outcome, as the overlap in predictions limits the additional value derived from human input.

Alternatively, we can infer from the second case in Proposition 2 that the centaur has a clear benefit over the ML algorithm when the human's predictions are accurate of the true outcome (i.e., $y_i - l(\mathbf{x}_i) \to 0$), and when they are dissimilar with the ML (i.e., $f(\mathbf{x}_i) - l(\mathbf{x}_i) \neq 0$). In this case, the benefit of having the human input will scale linearly with the variance of human predictions. This finding reflects that wider variations of the true outcomes corresponds to more variations in the human inputs (since they are assumed to be accurate representations of the true outcomes), and this, in turn, can yield a higher complimentary value in augmenting the ML model with human intuition.

Lastly, the third case in Proposition 2 reflects a more general insight for when human predictions are both dissimilar from the ML (i.e., $f(\mathbf{x}_i) - l(\mathbf{x}_i) \neq 0$) and the true outcome (i.e., $y_i - l(\mathbf{x}_i) \neq 0$). In such an event, we find that the benefit of the centaur over the ML depends greatly on (a) the correlation of the human's prediction and the actual outcome (i.e., $\rho_{\mathbf{y},l(\mathbf{X})}$), and (b) the baseline accuracy of the ML model (i.e., $\mathsf{MSE}(\mathbf{y}, f(\mathbf{X}))$). This result highlights two key insights regarding the role of human predictions in the centaur model. First, human predictions do not need to be highly accurate to add value. Put differently, the centaur model can extract and leverage complementary value from patterns or signals embedded in human predictions, even when they are imprecise, to improve upon standard ML algorithms. Second, the contribution of human predictions diminishes as other features in the data become more predictive. In such cases, the centaur model adapts by prioritizing the most informative inputs, shifting its reliance away from human predictions when the baseline ML model achieves higher accuracy. This adaptability underscores the centaur model's capacity to balance the strengths of human intuition and algorithmic precision, ensuring optimal use of available information.

Our findings from Proposition 2, particularly Case 3, highlight a potential three-way relationship between human predictions $l(\mathbf{X})$, machine learning predictions $f(\mathbf{X})$, and the true outcome $\mathbf{y}$. To elucidate this interaction, we conduct a sub-analysis in the electronic companion (see Proposition EC.1), which demonstrates that the benefit of incorporating human input is directly tied to the correlations among $f(\mathbf{X})$, $q(\mathbf{X}, l(\mathbf{X}))$, and $\mathbf{y}$. To build intuition, Figure 1 presents a simulation offering a visual representation of our findings. Specifically, we find that the benefit of human input diminishes as $l(\mathbf{X})$ becomes more correlated with $f(\mathbf{X})$, since the human's predictions no longer provide complementary information compared to the ML model. Conversely, when $l(\mathbf{X})$ is less correlated with $f(\mathbf{X})$, the human's input can significantly enhance the ML model's performance, provided there is a reasonably strong correlation between $l(\mathbf{X})$ and $\mathbf{y}$.

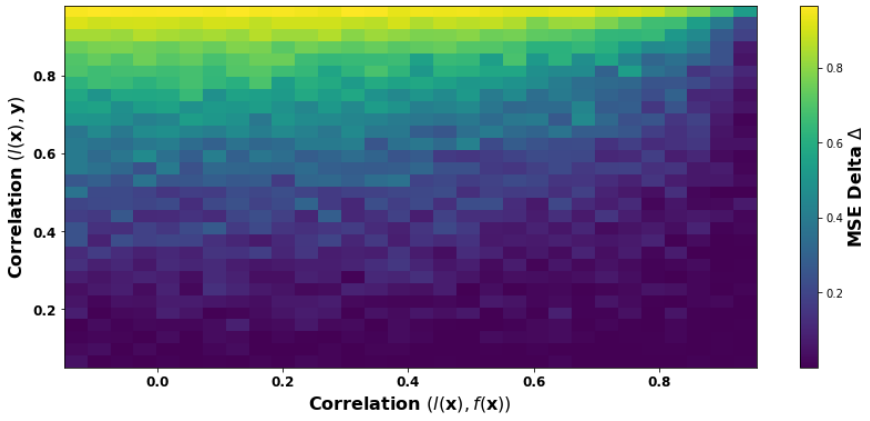


Figure 1 **Heatmap measuring Δ as $\rho_{\mathbf{y},l(\mathbf{X})}$, and $\rho_{l(\mathbf{X}),f(\mathbf{X})}$ are varied. The x-axis denotes $\rho_{l(\mathbf{X}),f(\mathbf{X})}$, the y-axis denotes $\rho_{\mathbf{y},l(\mathbf{X})}$. The values plotted represent Δ, the difference in MSE between the ML and Centaur.**

4. Operationalizing Centaurs in Practice

Building on our theoretical insights, we now present a generalizable framework for implementing centaur models in practice. Its prerequisite is a small-scale study (e.g., a survey) where human experts are required to provide their risk evaluation on the same task as the ML model. This study can be conducted using retrospective data alone, avoiding the need for significant changes or costly investments in IT infrastructure. Our framework offers a scalable, cost-effective approach to enhancing algorithmic performance without requiring continuous human supervision. Thus, practitioners could use our approach to evaluate the expected improvement that a centaur model could yield in their practice. We present a summary of the proposed framework in Figure 2 and describe each of the steps in greater detail as follows:

1. **Collect retrospective data:** The first step involves the compilation of the original dataset $(\mathbf{X}_N, \mathbf{y}_N)$ that will be used to derive the baseline ML model $f(\mathbf{X})$. Following the standard procedure for ML training and validation, we partition the data into a training set $(\mathbf{X}_T, \mathbf{y}_T)$ and testing set $(\mathbf{X}_E, \mathbf{y}_E)$ (see Section 5.1).
2. **Train a baseline ML model:** Using the observations $(\mathbf{X}_T, \mathbf{y}_T)$, derive the best possible ML model by training any supervised learning algorithm (see Section 5.2) and comparing their performance using the unseen samples of the entire testing set $(\mathbf{X}_E, \mathbf{y}_E)$.

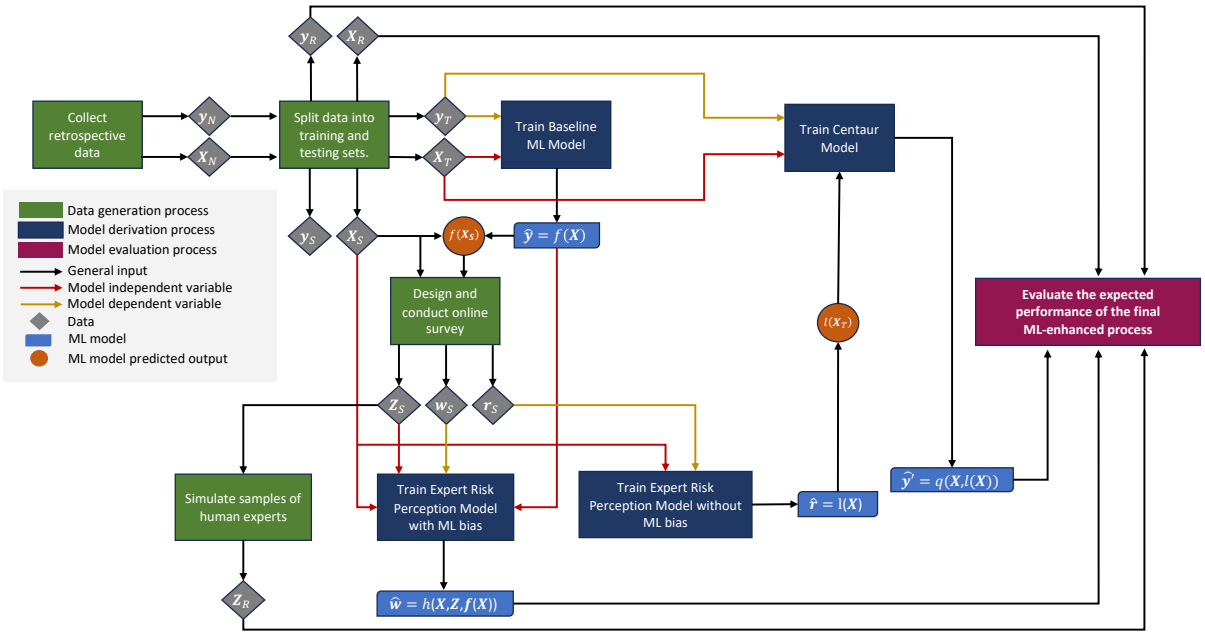


Figure 2 **Summary illustration of the generalizable centaur framework. We present in detail the flow of data, model training, validation, and evaluation that practitioners can follow to develop and deploy centaur models in practice.**

3. **Design and conduct an online survey:** Conduct a survey to collect human expert responses (see Section 5.3). Let $S \subset E$ be a subset of the testing set corresponding to the online survey. We use r_j, w_j, and $\mathbf{z}_j$ to denote the human expert's prediction without observing the ML predictions, the human expert's prediction after observing the ML predictions, and the vector of the human expert's characteristics that we record for each practitioner participating in the survey $j \in [J]$, respectively. Using as input the samples in $(\mathbf{X}_S, \mathbf{y}_S)$, along with the predictions of the baseline ML model $\hat{\mathbf{y}}_S$, the survey allows collecting $(\mathbf{Z}_S, \mathbf{w}_S, \mathbf{r}_S)$.
4. **Train expert risk perception models both without and with ML influence:** We derive a model $\hat{\mathbf{r}} = l(\mathbf{X})$ trained on the first series of expert responses $\mathbf{r}_S$ to capture systematically the human experts' perception of risk for the task of interest without being inform about the ML model's prediction (see Section 5.4). Similarly, we derive a second model $\hat{\mathbf{w}} = h(\mathbf{X}, \mathbf{Z}, f(\mathbf{X}))$ trained on the last series of survey responses $\mathbf{w}_S$ to codify the risk perception of human decision makers after being inform about the ML model's prediction while capturing the influence of the heterogeneity in the human experts' characteristics (see Section 5.4).
5. **Create the centaur model:** Using as input the expert risk perception model $l(\mathbf{X}_T)$ model with the original training set $(\mathbf{X}_T, \mathbf{y}_T)$, update the baseline ML model to derive the centaur model $q(\mathbf{X}, l(\mathbf{X}))$ (see Section 5.5).

The proposed centaur framework is an inexpensive way for organizations to incorporate the intuition of their own human experts into a baseline ML model. Practitioners can leverage this approach to not only improve the performance of ML models with human input but also estimate the ultimate impact that they will have upon deployment. An extension to our proposed centaur framework detailing how to conduct such an evaluation is described in Section EC.3. In Section 5.8 and 5.9 of our case study, we further demonstrate an application of this analysis for the context of transplantation readmissions predictions. The application of our framework shows that improvement can come at a low cost as the only additional required input is the responses from a retrospective survey.

5. Case Study: Predicting Solid Organ Transplantation Readmissions

In this section, we apply the centaur framework in a real-world case study in healthcare. We collaborate with Mayo Clinic Arizona to implement our model (using the steps described in the previous section) and make use of it to assist physicians in estimating the risk of 30-day readmission for solid organ transplantation. In Section 5.1, we describe the patient population clinical characteristics $\mathbf{X}$. Section 5.2 outlines the development and validation of our baseline ML model $f(\mathbf{X})$. Section 5.3 describes the survey conducted to capture the practitioner intuition and their characteristics $\mathbf{Z}$. Section 5.4 details the learned human risk perception model $l(\mathbf{X})$. Section 5.5 presents the centaur

model $q(\mathbf{X}, l(\mathbf{X}))$, and Section 5.6 compares the ML and the human risk assessment policies. In Section 5.7, we characterize the influence of ML models on human judgment using $g(\mathbf{X}, \mathbf{Z})$ and $h(\mathbf{X}, \mathbf{Z}, q(\mathbf{X}, l(\mathbf{X}))$. Lastly, in Sections 5.8 and 5.9, we estimate the impact of our model in practice via a simulation and an economic analysis.

5.1. Patient Data

Our analysis leverages retrospective clinical data obtained from electronic health records of Mayo Clinic Arizona's endocrinology and transplantation departments. Our data set comprises 1,537 de-identified cases of patients who received solid organ transplantation between September 25, 2015 and December 25, 2018. The study received IRB approvals from both Mayo Clinic and Harvard, and included only patients undergoing first-time solitary transplants. Table 1 presents a summary of all patient-specific variables utilized in the study (see Table EC.2 for a more detailed description). All information collected was available in the hospital's electronic health records at the time of each patient's discharge. We further incorporated organ-specific risk factors that we obtained from the United Network for Organ Sharing. 67.5% of the patients in our data set had kidney transplantation while 23.7% received a liver and 8.8% underwent heart transplantation. Overall, 23.0% of the patients in the study were re-admitted within 30 days from the index hospitalization. Missing information was imputed using the MedImpute algorithm to account for temporal data associations (Bertsimas et al. 2021).

5.2. The Baseline Machine Learning Algorithm

We trained multiple well-established ML algorithms to predict our outcome of interest (30-day readmission). We compared the performance of regularized logistic regression, with classification trees (CART), random forests (RF), gradient boosted trees (XGBoost), support vector machines

Category	Variables
Outcome	30-Day Readmission
Recipient Information	Age, Gender, Race, BMI, Length of Stay, History of Diabetes, Hyperglycemia, Hypoglycemia, HbA1c Value
Donor Information	Age, Gender, Race, BMI, Donor Type (Deceased or Living)
Transplant Factors	Organ Type, HLA Mismatch, Cold Ischemic Time, Graft Status, Delayed Graft Function, Time on Dialysis
Glucose Metrics	BG Average, BG Maximum, BG Minimum, BG Range, % BG Above 180, % BG Below 70
Metabolic Factors	Insulin Treatment (Basal/Bolus/None), Creatinine at Discharge
Liver-Specific Metrics	MELD Score, Bilirubin, Portal Vein Tumor Thrombus, Functional Status, Diagnosis (e.g., Cirrhosis Types)
Heart-Specific Metrics	LVAD Presence, Use of Inotropes, Functional Status, Diagnosis (e.g., Dilated Myopathy)
Kidney-Specific Metrics	EPTS at Transplant, Time on Dialysis, Graft Status

Table 1 Summary of variable categories and associated features for patient population.

(SVM), and multi-layer perceptron (MLP). To conduct unbiased tests in assessing the performance of these algorithms, we split the sample population into a training (75%) and a testing cohort (25%) for five stratified bootstrapped partitions of the data. We performed hyperparameter tuning using a bayesian optimization framework (Head et al. 2020) with the goal of maximizing the K-fold cross-validation area under the receiver operator curve (AUC).

Our results demonstrate that the XGBoost algorithm achieves superior performance (84.0% out-of-sample AUC) compared to the other methods considered (see Table EC.3). Our analysis suggests that out-of-sample AUC is higher (86.8%) for patients with a history of diabetes mellitus. We also observe variation in the ML algorithm's performance based on the type of organ. The mean AUC for kidney patients is 77.0%, but for liver cases, it reaches 97.5%, and for heart, it drops to 66.8%. Of note, the small sample size of the heart population (only 136 patients since heart transportation is a relatively rare operation) is the main reason behind this finding. Nevertheless, our ML algorithm still yields a better out-of-sample AUC compared to other widely used early readmission predictive methods that are applicable for heart transplantation patients (Sudhakar et al. 2015, Orfanoudaki et al. 2023). Finally, we find that combining cases across all solid organs significantly improved the predictive accuracy for liver samples even though they form 23.7% of the overall data set.

5.3. Survey Design

Following our proposed framework, we utilized an online survey platform to capture the risk perception of medical experts. We invited a diverse team of medical providers practicing at our partner hospital to respond to a series of questions about individual patient cases $(\mathbf{X}_S, \mathbf{y}_S)$ used to evaluate the baseline ML model. To ensure a fair comparison between the ML model and the respondents, both were provided exactly with the same information as captured in the patient feature vector $\mathbf{x}$. Each participant was tasked with reviewing up to five patient cases. Their objective was to analyze the provided patient data and submit responses to the questions outlined in Table 2. For each patient, the survey displayed all relevant case information available in the dataset through an interactive interface (see Figure EC.1). Questions were shown to the physicians in a sequential manner, and a response was required to allow the user to proceed to the next step. Once an answer was submitted, participants could not change their responses. Human experts were only informed regarding the ML prediction once reaching the last question (Q5), ensuring that (a) their initial risk assessments were not biased by the algorithmic recommendation, and (b) the survey could still capture the impact of the the algorithmic recommendation on human experts' risk perceptions.

For the first and fifth survey questions (Q1 and Q5), participants were specifically asked to estimate risk using intervals with 10% increments (e.g., $[0\%, 10\%), [10\%, 20\%)$, etc.). Prior research

Question	Details
Q1	What is the probability that the patient will require readmission within 30 days after discharge, according to your judgment?
Q2	What are the five most important clinical features that drove your decision among those listed here?
Q3	What would you change in the patient care during the index admission if you knew that the patient was at high risk upon discharge?
Q4	What other factors might contribute to patient readmission risk that are not listed here?
Q5	What do you think is the probability that the patient will require readmission within 30 days after discharge, after considering the ML model prediction?
Q6	What other factors might contribute to patient readmission risk that are not list here?

Table 2 Survey questions.

has shown that it is highly challenging for human assessors to differentiate between continuous values of risk (Sawyer 1966). Based on (Goldberg 1970), we avoided requiring participants to submit continuous risk values and instead assigned them to discrete yet granular categories. This approach reduced the cognitive load on participants and minimized the time needed to complete the survey, enabling the collection of a larger number of responses. Additionally, it allowed us to evaluate the discrimination performance of participants, providing deeper insights into expert risk perception.

Table 3 summarizes the survey responses. In total, 83 unique patients were reviewed with 125 distinct evaluations. 47.0% of the cases were reviewed by one expert and 53.0% by two. The average number of patients reviewed per respondent was 3.47, yielding a reasonable level of survey sample error equal to 8.0% (Dillman 2011).

5.4. Risk Perception Model of Human Experts

We next made use of our survey to derive a model of the human experts' intuition in predicting the risk of readmission. Specifically, we trained a ML model $l(\mathbf{X}) = \hat{\mathbf{r}}$ based on the survey responses $(\mathbf{X}_S, \mathbf{r}_S)$. We considered linear regression, CART, RF, XGBoost, and SVM as candidate models for this task. Table EC.4 summarizes the predictive performance of the ML methods considered. We found the linear model as the best to capture the human experts' responses, with a mean absolute error (MAE) of 0.111 and a Brier Score of 0.020. This finding is consistent with numerous studies from the psychology literature (Karelaia and Hogarth 2008, Cooksey 1996). Non-linear models

Metric	Value
Total number of respondents	38
% of Doctors of Medicine (Advanced Practice Providers)	68.42% (31.58%)
% specializing in transplantation (endocrinology)	31.58% (68.42%)
Mean years of professional experience	17.26
Standard deviation of years of professional experience	10.94
Average number of patient records reviewed per expert	3.47

Table 3 Survey participant response details.

overfit the responses in the survey, obtaining low predictive power. In addition to the superior predictive performance, the ordinary least squares model was also attractive because it allowed us to perform inference and identify the influence of the primary independent variables that drive human risk assessments. We note that such a model is necessary for our study due to the lack of retrospective data on the practitioner's predicted readmission risk for every patient. Thus, we develop a model to efficiently obtain an estimate of human readmission risk perception for all $1,537$ patients in our dataset.

For our final model, we only included independent variables with significant t-tests in our model. We further examined the residuals for linearity, heteroscedasticity, auto-correlation, and outliers. The ordinary-least squares (OLS) regression coefficients, along with the resulting $p-$values of the $t-$tests, are summarized in Table EC.5. The expected value of the predicted risk is $\sum_{i \in T} \frac{1}{|T|} l(\mathbf{x_i}) = 23.0\%$ with a 18.0% standard deviation. The learned $l(\mathbf{X})$ is by design only a function of the patient characteristics and excludes physician characteristics, allowing us to directly apply this learner to all patient observations in our dataset (even those that were not part of the survey) and perform a fair comparison between the original ML model and the centaur. We present our results related to the inclusion of the human experts' characteristics $\mathbf{Z}$ in Section 5.7.

5.5. The Centaur Model

In this section, we examine the performance of our centaur model to predict readmission risk for transplant patients. Specifically, we augment the patient features $\mathbf{X}$ with the outputs from the model $l(\mathbf{X})$ developed in 5.4 to form a new feature vector $\{\mathbf{X}, l(\mathbf{X})\}$. This new set of features is then used by a revised XGBoost model to form our centaur $q(\mathbf{X}, l(\mathbf{X}))$. Table 4 presents a comparison of the centaur to the baseline ML and human expert perception models.

We find that the centaur ($q(\mathbf{X}_R, l(\mathbf{X}_R))$) outperforms both the standalone ML model ($f(\mathbf{X}_R)$) and the human expert risk perception model ($l(\mathbf{X}_R)$) in terms of out-of-sample AUC, accuracy, and correlation with the true outcome. The centaur achieves an out-of-sample AUC improvement of 2.42% in absolute value (equivalent to a 2.89% relative improvement) over the best ML model. These values are considerable given that best ML model already represents a high-performing baseline (AUC = 84%), improvements beyond which are exponentially hard. Our results demonstrate that the human predictions ($l(\mathbf{X}_R)$), despite achieving a relatively low standalone AUC of 57.11%

Model	AUC	Accuracy	Correlation with $\mathbf{y}_r$
Human Expert Risk Perception Model ($l(\mathbf{X}_R)$)	57.11%	75.58%	0.0794
Best ML Model ($f(\mathbf{X}_R)$)	84.00%	86.75%	0.5791
Centaur Model ($q(\mathbf{X}_R, l(\mathbf{X}_R))$)	86.42%	88.05%	0.6276

Table 4 **Comparison of $q(\mathbf{X}_R, l(\mathbf{X}_R))$, $f(\mathbf{X}_R)$, and $l(\mathbf{X}_R)$ on the testing set $(\mathbf{y}_r, \mathbf{X}_R)$.**

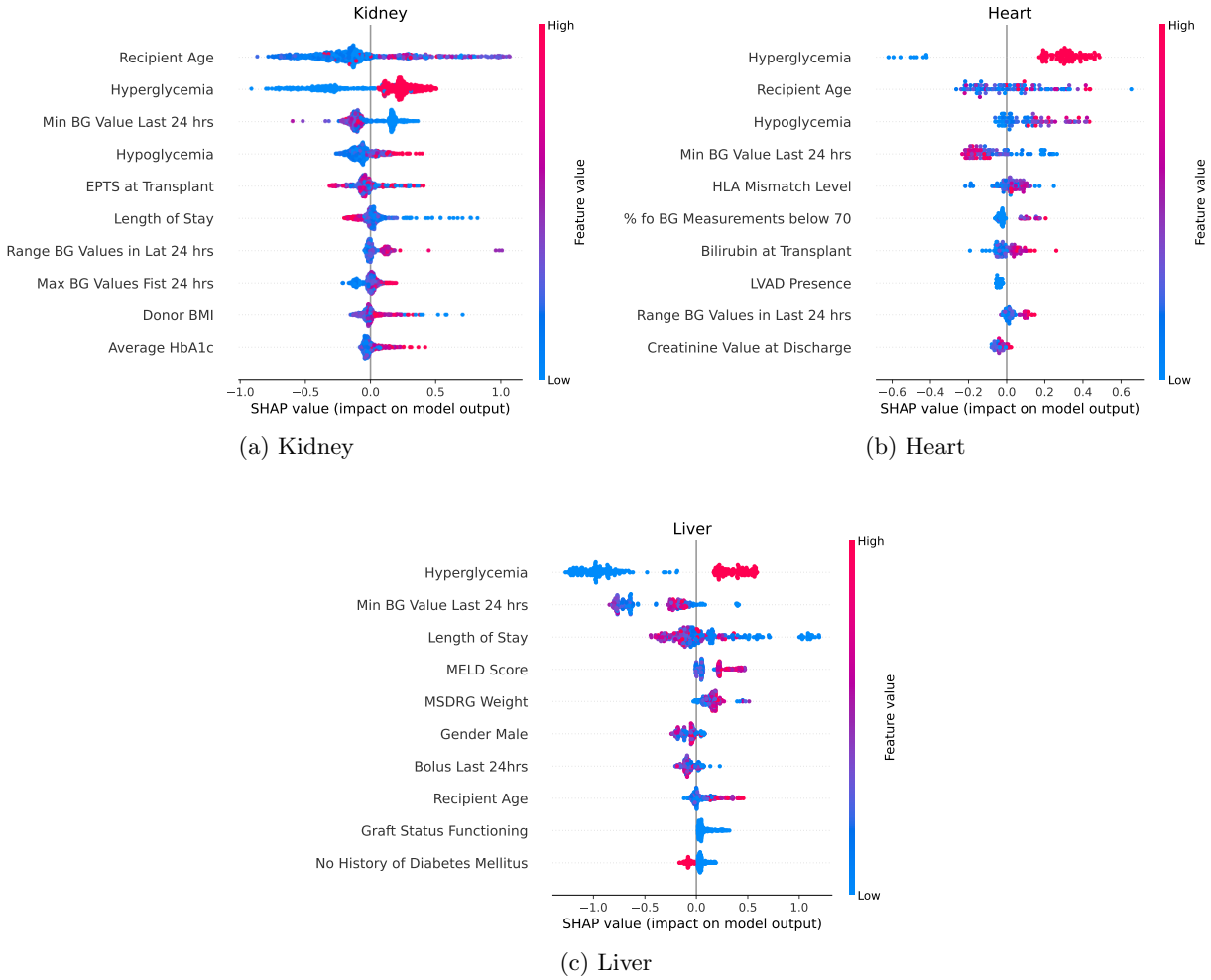
and a modest correlation with the true outcome, contribute unique and complementary information that enhances the overall predictive performance of the centaur model. The low correlation between the human expert model ($l(\mathbf{X}_R)$) and the best ML model ($f(\mathbf{X}_R)$), quantified at 0.01517, underscores the distinct nature of the information contributed by each source. This is consistent with our theoretical insight that human input, even when suboptimal, can improve model performance if it provides information not captured by the ML model. Put differently, the centaur model realizes that although human assessments are, on average, worse than the ML model, for some specific cases they might be superior. Thus, by adaptively weighting human assessments depending on the patient characteristics, the centaur provides a meaningful way of incorporating human intuition into the ML model.

5.6. Comparative Analysis of Human and Algorithmic Performance

Proposition 2 highlights that integrating human risk perceptions into the ML model to create a centaur can enhance predictive performance when human and ML risk estimation policies diverge. To validate this theoretical insight, in this section, we analyze how human and ML risk assessments differ in their reliance on independent variables and in the magnitude of their predictions.

ML versus Human Reasoning. First, we use the SHapley Additive exPlanations (SHAP) framework (Lundberg et al. 2020) to derive clinically relevant insights for medical practitioners from the baseline ML model ($f(\mathbf{X})$) and identify the key independent variables predicting early hospital readmission for each organ type. These variables are then compared to those prioritized by human experts in their risk assessments. Figure 3 displays the 10 most important features for each organ based on the best ML model (XGBoost), ranked by significance. Higher feature values are shown in red and lower values in blue. Positive (negative) SHAP values indicate a greater (lower) likelihood of 30-day readmission.

The second survey question (Q2) asked participants to specify five key patient characteristics driving their risk estimations. Their responses suggest that human intuition favors summary metrics and past comorbidities over measures of variability, aligning with past studies on clinical risk drivers after transplants (see Section 2). In the electronic companion, Figure EC.2 summarizes the human expert responses by organ type, with corresponding p-values in Table EC.8. Our results highlight that experts frequently identified the same set of factors as primary drivers across cases, regardless of the patient's condition. However, inter-rater agreement measured using Fleiss Kappa (Fleiss 1971) was low. For instance, in over 50% of cases, human expert participants indicated diabetes history among the top five factors, yet the agreement among expert reviews of the same case was limited ($k = 0.094$). These results indicate that while experts tend to select similar risk factors, their evaluations vary significantly when reviewing the same patient (see Appendix EC.9).



(a) Kidney (b) Heart

(c) Liver

Figure 3 **SHAP Plots for the proposed XGBoost ML model summarizing the risk contribution of the ten most important features per organ type. Acronyms are defined in the notes of Table EC.2.**

More importantly, we found a stark difference between the key independent variables that influence the estimations of the ML model and the human experts. The ML model places a lot of emphasis on various BG metrics during hospital admission, including the minimum and maximum values and the presence of hyperglycemia and hypoglycemia. These metrics capture the variability of a patient's metabolic condition throughout the hospital stay. In contrast, medical experts identified across all organs the mean BG value and history of diabetes as two of the most important determinants of readmission risk. The judgment of the human experts was also driven to a higher degree by organ-specific variables, including the presence of LVAD (heart), the organ's functional status at transplant (liver), and listing (heart). Finally, we found that in some organs (e.g., Kidney) some variables (e.g., race) play a role in human expert assessments, but not in the ML model.

ML versus Human Risk Misperception. We next further investigated the misperception of risk among the ML model and the human expert responses. In particular, we analyzed the human

expert predictions' degree of divergence from the ML model. This, in turn, enabled us to uncover the settings in which human experts' intuition is more accurate than the ML algorithm, justifying the added benefit of the proposed centaur approach.

Table 5 shows the proportion of cases where human experts over- or underestimated the risk of readmission compared to the ML model. In the first question (Q1), where providers had no access to the ML estimation, 4.17% (14.5%) of participants agreed with the algorithm's recommendation for cases of readmission (no readmission). When experts overestimated the risk, 62.38% of the cases were not associated with a readmission. These findings align with the low out-of-sample AUC of the provider's estimations (see Section 5.7). After being prompted with the ML output in the fifth question (Q5), agreement rates improved significantly to 22.77% (8.33%) for no readmission (readmission), while the frequency of under- and overestimation decreased as providers better calibrated their responses. Notably, in 33.33% of readmission cases, human experts outperformed the algorithm by predicting higher risk more accurately. This highlights the potential value of human insights in alignment with our theoretical findings. To further explore this, Figure EC.3 examines the average predicted risk across provider subgroups. Two key insights emerge: (1) human experts, on average, overestimate the underlying risk (i.e., adopt a more conservative stance) compared to the ML algorithm, and (2) this overestimation behavior is consistent across subgroups.

5.7. The Influence of the ML Recommendations on Human Experts' Assessments

In this section, we investigate the impact of the ML recommendations on human experts' assessments. Following our framework (see Section 4), we train two separate OLS models using response of the medical practitioners to the first and fifth survey questions (Q1 and Q5, respectively). Both models use as independent variables the patient information $\mathbf{X}$ and the expert's characteristics $\mathbf{Z}$.

Using the responses to Q1, we construct the model $g(\mathbf{X}, \mathbf{Z})$, representing human risk assessments without knowing the ML prediction. Our model development first removed all the independent variables with insignificant t-tests. With the reduced model, we examined the residuals for linearity,

Outcome	Human Risk=ML Risk	Human Risk<ML Risk	Human Risk>ML Risk
Participant responses before receiving the ML estimation (Q1)			
No Readmission	14.85%	22.77%	62.38%
Readmission	4.17%	66.67%	29.17%
Participant responses after receiving the ML estimation (Q5)			
No Readmission	22.77%	21.78%	55.45%
Readmission	8.33%	58.33%	33.33%

Notes. The table summarizes the proportion of cases where experts under-predicted, over-predicted, or were in agreement with the algorithm's estimations.

Table 5 Comparison of algorithm and human estimations before and after the introduction of the ML model in the online survey.

heteroscedasticity, auto-correlation, and outliers. Due to the limited sample size, we trained the model on the entire population of survey responses (125 observations). The MAE of the Q1 model ($g(\mathbf{X},\mathbf{Z}) = \hat{\mathbf{r}}$) was 0.1075 and the associated Brier score was 0.0204.

Using Q5, we assess the impact of the ML recommendation, $f(\mathbf{X})$, by including the ML prediction as an additional independent variable to form the model $h(\mathbf{X},\mathbf{Z},f(\mathbf{X}))$, representing human risk assessments after knowing the ML prediction. This model accounts for both the direct influence of the ML model and changes in the relative importance of other variables describing patient and practitioner characteristics. The MAE of the Q5 model was 0.1146 and the Brier score was 0.0252. The linear regression coefficients for both models, along with the resulting $p-$values of the $t-$tests, are summarized in Table 6.

Next, we evaluate how different groups of medical experts responded to ML recommendations (30-day readmission risk prediction in our setting) and assess the resulting impact on their predictive accuracy. We measure the degree of algorithm influence using the weight of advice (WoA) metric (Harvey and Fischer 1997):

$$\text{WoA} = \frac{\text{final expert estimate} \;-\; \text{initial expert estimate}}{\text{ML algorithm estimation} - \text{initial expert estimate}}.$$

Higher values indicate that the decision-maker significantly relies on the algorithm's advice, while a value of 0 signifies that the decision-maker completely ignores the advice. We exclude from the metric all cases where the initial human estimate matched the algorithm's recommendation. The WoA metric reflects the degree to which clinicians weigh the algorithm's advice and so is inversely related to algorithm aversion and discounting.

Figure EC.4 illustrates that ML recommendations positively influenced all groups, irrespective of baseline accuracy. Table 7 summarizes our results related to performance of subgroup of experts

Regression Model	$g(\mathbf{X},\mathbf{Z}) = \hat{\mathbf{r}}$		$h(\mathbf{X},\mathbf{Z},f(\mathbf{X})) = \hat{\mathbf{w}}$		$l(\mathbf{X})$	
Independent Variable	*OLS Coefficient*	*p-value*	*OLS Coefficient*	*p-value*	*OLS Coefficient*	*p-value*
Constant	-0.6169	<0.001	-0.3113	0.066	-0.5012	0.002
Patient Information						
Recipient Age at Admission	0.0029	0.012	0.0028	0.020	0.0023	0.043
Recipient BMI	0.0083	0.004	0.0022	0.0472	0.0081	0.006
Creatinine Value at Discharge	0.0071	0.023	0.0044	0.0475	0.0033	0.039
Average BG Value in Last 24 hrs	0.0017	<0.001	0.0007	0.0151	0.0015	0.001
History of Diabetes 2 Mellitus	0.0124	0.008	0.0150	0.0763	0.0161	0.0073
HbA1c at Admission	0.0285	0.004	0.0235	0.0127	0.0246	0.0101
Human Expert Information						
Role: MD	-0.0673	0.032	-0.0719	0.028	n/a	n/a
Years of Professional Experience	0.0041	0.013	0.0029	0.095	n/a	n/a
ML Recommendation	n/a	n/a	0.2533	0.003	n/a	n/a

Table 6 Output summary of the linear regression models capturing the human experts' risk perception. We report the resulting coefficients only for the reduced models and the associated p-values of the t-tests.

Clinical Subgroup	Experts AUC without ML	Experts AUC with ML	WoA	Improvement
All participants	55.03%	61.24%	36.33% (14.4%)	11.28%
Transplantation	64.82%	87.68%	54.12% (14.89%)	35.26%
Endocrinology	50.87%	52.22%	25.71% (14.1%)	2.65%
doctors of medicine (MDs)	59.28%	64.35%	43.04% (14.67%)	8.55%
advanced practice provider (APPs)	48.34%	56.48%	26.36% (14.0%)	16.84%
Experience $\leq$ 12 years	57.99%	65.45%	37.64% (12.0%)	12.85%
Experience $\geq$ 12 years	53.61%	58.89%	35.42% (16.0%)	9.84%

Notes. We report the resulting AUC metrics for the responses provided both before (Q1) and after (Q5) the introduction of the ML model's recommendations per expert subgroup. The table includes the WoA metric for all participant groups considered. In parenthesis, we indicate the percentage of responses in which the first response of the human expert matched the ML recommendation. The last column measures the % relative improvement of physicians' estimation AUC with the help of ML.

Table 7 Discrimination performance summary of clinical experts' evaluations on the task of 30-day readmission.

both with and without ML recommendations as well as their WoA measure and gained improvements. Overall, we observe that the ML estimations positively influenced the survey participants across all clinical subgroups. Overall, transplantation experts achieved higher discrimination performance (64.82%) than their endocrinology peers (50.87%). Similarly, MDs outperformed APPs, and practitioners with more than 12 years of experience exhibited slightly higher AUC scores. Statistical tests confirmed that all observed differences in AUC between the ML model and expert groups, with and without algorithmic recommendations, were highly significant ($p < 0.001$). The WoA results reveal that transplantation experts and MDs exhibited the highest reliance on ML recommendations, with transplantation experts showing the greatest relative improvement. Although APPs had lower WoA scores, they achieved twice the relative improvement in AUC compared to MDs. No significant differences in WoA were observed across experience levels.

5.8. Measuring the Impact of the Centaur in Practice

In this section, we illustrate the application of the proposed generalizable framework to our study, illustrating the last step presented in Section 4. First, we simulate a sample of human experts using the synthetic data vault package (Patki et al. 2016). For each patient i in the external testing set R, we generate an MD $\mathbf{Z}_{MD,i}$ and an APP $\mathbf{Z}_{APP,i}$ with characteristics captured by the vector $\mathbf{Z}$ who are called to independently assess their risk of transplantation. We summarize the estimated AUC performance of the derived models on the external testing set R in Table 8. The AUC of the original baseline ML and the centaur models on this sample population are $AUC(f(\mathbf{X}_R), \mathbf{y}_R) = 81.61\%$ and $AUC(q(\mathbf{X}_R, l(\mathbf{X}_R)), \mathbf{y}_R)) = 83.61\%$ respectively. As described above, we use the $h(\mathbf{X}, \mathbf{Z}, q(\mathbf{X}, l(\mathbf{X}))$ model to estimate the downstream performance of the ML-enhanced process that the humans will follow. In our setting, we independently approximate the performance of the MDs and the APPs. We find that the performance of the MDs is higher ($AUC(h(\mathbf{X}_R, \mathbf{Z}_{MD}, q(\mathbf{X}_R, l(\mathbf{X}_R)), \mathbf{y}) = 73.54\%$)

Model	Description	$[\mathbf{X}_R, \mathbf{Z}_{MD}]$	$[\mathbf{X}_R, \mathbf{Z}_{APP}]$
$g(\mathbf{X}, \mathbf{Z})$	Expert risk perception model	55.42%	50.76%
$h(\mathbf{X}, \mathbf{Z}, q(\mathbf{X}, l(\mathbf{X})))$	Centaur-enhanced expert risk perception model	73.54%	71.83%

Table 8 Summary of AUC performance of the risk estimation models developed and validated as part of the centaur framework. The metric is evaluated against the ground truth labels $\mathbf{y}_R$ across all models.

compared to the nurses ($AUC(h(\mathbf{X}_R, \mathbf{Z}_{MD}, q(\mathbf{X}_R, l(\mathbf{X}_R)), \mathbf{y}) = 71.83\%$). We perform the same evaluations using the $g(\mathbf{X}, \mathbf{Z})$ risk perception model which approximate the perception of the human decision makers without the bias of any ML suggestion. We find that $AUC(g(\mathbf{X}_R, \mathbf{Z}_{MD}), \mathbf{y}) = 55.42\%$ and $AUC(g(\mathbf{X}_R, \mathbf{Z}_{APP}), \mathbf{y}) = 50.76\%$.

Our analysis provides a well-grounded approximation of the final predictive accuracy of the centaur-enhanced expert risk estimation process. We demonstrate that the deployment of the centaur model can significantly improve the humans' risk assessment. The benefit of the centaur depends on the predictive accuracy of the human experts involved in the study as well as the capability of the model used to capture their predictions. There might be better ways to further improve performance by following a different way of incorporating human expertise into the algorithm. We leave it to future research to further investigate this, and thereby, develop even stronger centaurs for implementation in practice.

5.9. The Economic Value of the Centaur

We now answer the following question: what is the economic value that our centaur approach could bring to the healthcare system when used in our application? Leveraging the sixth survey question, we asked survey respondents what actions they would pursue if they knew that a patient would require readmission. Table EC.9 summarizes the responses we received. From our questionnaire, we identify seven primary categories of action and develop a simulated model of the expected economic value of a 30-day risk assessment estimator.

To estimate the economic value, we let $\tau \in [0, 1]$ be the risk threshold for administering interventions at our partner hospital, with an average intervention cost of WC and a readmission cost of RC. Interventions are not fully effective, with only a proportion p of treated patients avoiding readmission. Under these assumptions, the annual economic value of a readmission prediction model m is given by:

$$EV_m = p \cdot RC \cdot RP_{\tau,m} - WC \cdot P_{\tau,m},$$

where $RP_{\tau,m}$ is the expected annual number of patients with predicted risk exceeding τ who are readmitted despite intervention, and $P_{\tau,m}$ is the total annual number of patients exceeding τ under model m. We perform a sensitivity analysis on p and WC, setting $RC = \$27,000$ (Weiss and Jiang 2021). We compare the economic performance of the human-only risk model (current

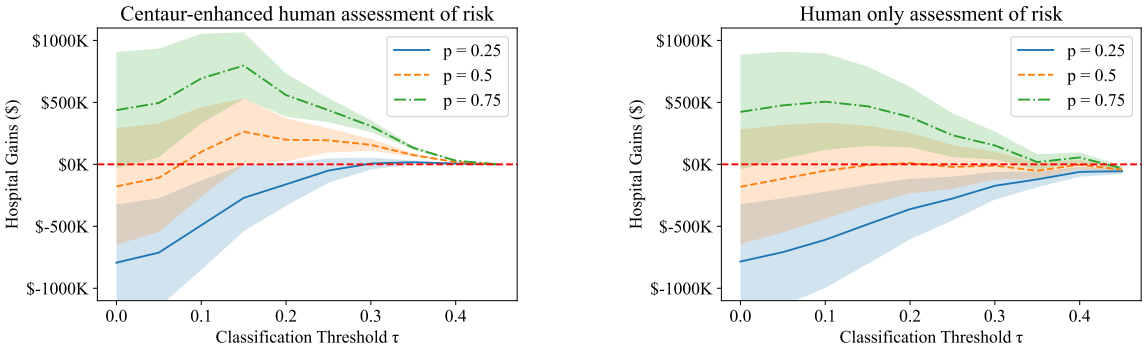


Figure 4 **Comparison of estimated economic gains for the Mayo Clinic between the existing process of readmission risk evaluation that involves only human judgment and the proposed centaur-enhanced process where medical practitioners receive recommendations from the proposed centaur model.**

practice) and the centaur-enhanced process across risk thresholds $\tau \in (0, 0.45]$. Intervention costs range from $WC \in [2{,}000,\ 4{,}000]$ (Mann et al. 2020, Serrano et al. 2019), and intervention success probabilities are $p \in 0.25, 0.5, 0.75$. Figure 4 summarizes the potential annual cost savings of the centaur-enhanced approach at our partner hospital.

Our results indicate that the centaur approach is expected to offer significantly higher cost savings compared with the current practice in which human experts use their own risk predictions. When providers do not have access to the centaur predictions (i.e., the current practice), the hospital is incurring losses due to readmissions across all potential classification thresholds for lower values of p. On the contrary, a centaur-enhanced assessment can lead to significant gains even when only 50% of the interventions are effective. In addition, the centaur allows the hospital to reduce the number of ineffective interventions for the same range of classification thresholds ($\tau \in [0.1, 0.2]$), allowing the system to administer preventive care to patients who are truly in need. Finally, in the human-only process, our results indicate that the hospital is unable to gain benefits with more conservative classification thresholds due to the high number of false positive cases. Overall, since developing and testing the centaur model is fairly inexpensive, we find that it can offer sufficient financial incentives to hospital administrators for implementation in clinical practices.

6. Managerial Insights

We now synthesize the key insights derived from our study and outline critical considerations for decision-makers aiming to implement a centaur model effectively in practice.

6.1. Algorithm or Centaur?

Our findings provide empirical evidence that creating centaurs by incorporating human experts' insights in the form of a systematic model can substantially improve performance of pure ML algorithms. Our theoretical analysis in Section 3 highlights that the power of the centaur comes

from the complementarity between human intuition and ML power. Specifically, we find that the benefit of incorporating human input diminishes as human and ML predictions align more closely, regardless of the human's prediction accuracy. This suggests that when human and algorithmic predictions converge, the additional effort to include human judgment may not justify the cost. Second, human predictions are most valuable when they are accurate relative to the true outcomes and differ meaningfully from the ML predictions, particularly in scenarios with high variability in outcomes. Lastly, human predictions do not need to be highly accurate to provide value, as they can convey inherent insights into relationships through subtle signals, such as patterns of correlation.

Our numerical experiments further support the finding that our proposed centaur can improve the downstream performance of the algorithm even when the set of independent variables ($\mathbf{X}$) does not change. While the models $g(\mathbf{X}, \mathbf{Z})$ and $l(\mathbf{X})$ can accurately capture the experts' responses $\mathbf{r}$, their performance on the task of estimating the risk of readmission $\mathbf{y}$ is still poor (AUC $< 60\%$). This result suggests that humans are consistently predicting a dependent variable $\mathbf{r}$ which represents the perception of readmission risk. Still, this variable significantly differs from the target variable $\mathbf{y}$. The centaur model is able to identify cases where human input improves upon ML predictions. Figure EC.5 visualizes the differences in overall proportion of algorithm and physician responses in each risk category. Notably, while the expert AUC was, in general, low, we observe that there are cases where humans are better compared to the ML model. This is partially why the centaur model performs better than both the pure ML model and the human experts: the centaur model benefits from this heterogeneity, enabling it to effectively capture the complementary value of human intuition by relying more on human intuition when human intuition is valuable and less when it is not. Due to its design and structure, the centaur can systematically capture the cases in which the human risk perception is beneficial and leverage it to its advantage.

6.2. Human Perception

Our findings provide evidence that humans do not tend to place as much emphasis on metrics that capture fluctuation and variability but instead focus on summary metrics, such as the expected value or past comorbidities. Clinical intuition, as manifested in the providers' responses, is in line with the studies on clinical drivers of risk after a transplant that were outlined in Section 2. We also observe that physicians were very likely to highlight the same set of factors as the primary drivers of their judgment independent of the patient's depicted condition. However, when we used Fleiss Kappa statistic (Fleiss 1971) to measure the inter-rater agreement between participants that reviewed the same patient, we observed a high degree of consensus ($k > 0.2$) for only a small subset of variables (see Table EC.7). For example, in more than 50% of the patient cases reviewed, "presence of history of diabetes" was among the top five most important variables selected by

the participants. Nevertheless, we observed limited agreement (with respect to readmission risk) between providers that reviewed the same case ($k = 0.094$). These findings suggest that, while human experts based their judgment on the same risk factors, they were not necessarily in a high degree of agreement when reviewing the same patient.

From our results, we also find that when provided with exactly the same information, human experts are less accurate compared to the ML algorithm. At the same time, our analysis reveals that providing the algorithm's estimation as an input to clinicians at the time of the decision can positively influence their perceptions of risk. The degree of improvement depends on the confidence and WoA that human decision-makers place on the model. Thus, our survey highlights that algorithm aversion is a key barrier for human experts to achieve better performance.

6.3. Insights for Medical Practitioners

Our analysis demonstrates that glucometrics are highly predictive of early hospital readmission following solid organ transplantation, with blood glucose control during the initial hospital stay proving more indicative of future patient trajectories than a history of diabetes. Both high and abnormally low blood glucose levels (e.g., hypoglycemia and minimum glucose values during the last 24 hours) are linked to increased readmission risk. Additionally, a wider glucose range during the final 24 hours of the index admission is associated with higher readmission probability for kidney and heart patients. While diabetes history is a known risk factor, it ranks as a top feature only for liver transplants, underscoring the critical importance of real-time glucose management. To the best of our knowledge, this is the first study to establish a strong association between glucometrics during index admission and early hospital readmission across all major solid organ types. While Orfanoudaki et al. (2023) first highlighted the impact of suboptimal glucose metrics on readmission for kidney transplants, our findings generalize this relationship to heart and liver transplants as well. These insights offer actionable strategies for improving post-transplant care, emphasizing the need for stringent glucose control across diverse patient populations.

6.4. Economic Value Survey

We also assessed the impact that the proposed ML model could have on patient care by asking practitioners what actions they would pursue if they knew that a patient would require readmission. We identified seven primary categories of action whose frequency varies depending on the role and specialty of the provider.

Our results show that in 40% of cases, providers would not change patient care, with APPs reporting a lower rate (30%) compared to MDs (46.67%). The most common action taken was improving glycemic control, identified by 24% of respondents as a way to prevent re-hospitalization,

particularly for diabetic patients. While endocrinology and transplantation teams showed no significant differences in this regard, APPs were more likely than MDs to emphasize glycemic control (32% vs. 18.67%). APPs and the endocrinology team also prioritized treatment education to ensure adherence to post-transplantation and metabolic therapy, whereas MDs focused on close patient monitoring, such as extending hospital stays, scheduling early follow-ups, and monitoring organ health. Transplantation physicians further highlighted the importance of caregiver support at home, particularly for elderly patients, to prevent early readmissions.

This analysis highlights key clinical and operational strategies that transplantation centers could adopt to improve patient outcomes and reduce readmissions. It demonstrates the willingness of clinical teams to adapt their practices based on the services they provide. As healthcare systems transition to value-based care and higher levels of quality transparency (Saghafian and Hopp 2019), the effectiveness of ML-generated risk scores hinges on their ability to drive actionable changes in patient care for high-risk individuals. Integrating such models into systems like electronic health records could help flag high-risk patients at discharge. Providers could then implement interventions in five key areas: reviewing glycemic treatment, enhancing patient education, scheduling early follow-ups, extending hospital stays, and arranging home care support. These processes would tie ML risk assessments to actionable care decisions, improving patient pathways and outcomes.

7. Conclusions

Our research introduces a human-algorithm centaur model that systematically integrates human intuition with ML algorithms to enhance predictive performance in high-stakes decision-making. Theoretically, we establish that the centaur model outperforms standalone human or ML approaches by leveraging the unique, complementary insights humans provide, particularly when their predictions capture information orthogonal to the ML model. Empirically, our research provides evidence that algorithms can be more accurate than humans in predicting 30-day readmissions for organ transplant patients. Our survey reveals that ML algorithms can positively influence human experts' perception of risk depending on the degree of algorithm aversion. We find that clinicians often pay attention to different risk factors compared to algorithms. We also show that the centaur outperforms both the best pure ML predictions and those of human experts. Finally, we propose a generalizable approach to develop and validate human-algorithm centaur models, allowing practitioners to estimate the actual improvement in predictive capability and economic value of implementing them.

There are several limitations to the case study. First, the results are based on a retrospective analysis, leveraging data from a single medical center. Second, the comparison between the human experts and the centaur does not consider human tacit knowledge, which is not codified in structured variables, but can affect clinical decisions during medical care at the hospital. Third, we

focused on the 30-day readmission rate, which is the most common way of measuring patient returns used by various organizations including the Centers for Medicare & Medicaid Services. We did not consider longer-term horizons of potential readmission, such as a 60-day or 90-day window. Furthermore, survey participants highlighted that there are other factors outside our data that could play a role in the decisions made for patients, including quality of care, adherence to medication, and the socioeconomic background of the organ recipient. We leave it to future research to extend our analyses by collecting data on such factors. Finally, we found that human experts on average over-estimate risks. However, there might be many reasons why human-experts tend to over- or under-estimate risk compared to the ML model. Studying what is driving the physician expert behavior is beyond our area of focus, but presents an interesting extension to this work.

Notwithstanding these limitations, our study provides a systematic paradigm for modern organizations to develop centaur models that augment both human and algorithmic decision-making. We believe that our work provides a useful step towards this goal, as it generates important insights into how the power of algorithms and human intuition can be combined in high-stake decision-making settings such as those in care delivery for transplant patients. Given the power of human-algorithm centaurs, it is not illogical to think that one potential path for the future development and implementation of ML and AI algorithms is centaur-enhanced. Thus, we hope to see more studies centered around understanding and improving human-algorithm centaur models.

References

Arnold J, Davis A, Fischhoff B, Yecies E, Grace J, Klobuka A, Mohan D, Hanmer J (2019) Comparing the predictive ability of a commercial artificial intelligence early warning system with physician judgement for clinical deterioration in hospitalised general internal medicine patients: a prospective observational study. *BMJ open* 9(10):e032187.

Athey SC, Bryan KA, Gans JS (2020) The allocation of decision authority to human and artificial intelligence. *AEA Papers and Proceedings*, volume 110, 80–84 (American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203).

Bachmann JM, Shah AS, Duncan MS, Greevy Jr RA, Graves AJ, Ni S, Ooi HH, Wang TJ, Thomas RJ, Whooley MA, et al. (2018) Cardiac rehabilitation and readmissions after heart transplantation. *The Journal of Heart and Lung Transplantation* 37(4):467–476.

Bailey PE, Leon T, Ebner NC, Moustafa AA, Weidemann G (2022) A meta-analysis of the weight of advice in decision-making. *Current Psychology* 1–26.

Balakrishnan M, Ferreira K, Tong J (2022) Improving human-algorithm collaboration: Causes and mitigation of over-and under-adherence. *Available at SSRN 4298669* .

Beam AL, Kohane IS (2018) Big data and machine learning in health care. *JAMA* 319(13):1317–1318.

Beraha M, Metelli AM, Papini M, Tirinzoni A, Restelli M (2019) Feature selection via mutual information: New theoretical insights. *2019 international joint conference on neural networks (IJCNN)*, 1–9 (IEEE).

Bertsimas D, Orfanoudaki A (2021) Algorithmic insurance. *arXiv preprint arXiv:2106.00839* .

Bertsimas D, Orfanoudaki A, Pawlowski C (2021) Imputation of clinical covariates in time series. *Machine Learning* 110(1):185–248.

Boloori A, Saghafian S, Chakkera HA, Cook CB (2015) Characterization of remitting and relapsing hyperglycemia in post-renal-transplant recipients. *PLoS One* 10(11):e0142363.

Boyacı T, Canyakmaz C, de Véricourt F (2024) Human and machine: The impact of machine input on decision making under cognitive limitations. *Management Science* 70(2):1258–1275.

Breiman L (1996) Stacked regressions. *Machine learning* 24:49–64.

Caro F, de Tejada Cuenca AS (2023) Believing in analytics: Managers' adherence to price recommendations from a dss. *Manufacturing & Service Operations Management* 25(2):524–542.

Chakkera HA, Weil EJ, Castro J, Heilman RL, Reddy KS, Mazur MJ, Hamawi K, Mulligan DC, Moss AA, Mekeel KL, et al. (2009) Hyperglycemia during the immediate period after kidney transplantation. *Clinical Journal of the American Society of Nephrology* 4(4):853–859.

Chen CH (1976) On information and distance measures, error bounds, and feature selection. *Information Sciences* 10(2):159–173.

Choudhary V, Marchetti A, Shrestha YR, Puranam P (2023) Human-ai ensembles: When can they work? *Journal of Management* 01492063231194968.

Cooksey RW (1996) *Judgment analysis: Theory, methods, and applications.* (Academic press).

Dai T, Singh S (2021) Artificial intelligence on call: The physician's decision of whether to use AI in clinical practice. *Available at SSRN* URL `http://dx.doi.org/10.2139/ssrn.3987454`.

Dietvorst BJ, Simmons JP, Massey C (2015) Algorithm aversion: people erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General* 144(1):114.

Dillman DA (2011) *Mail and Internet surveys: The tailored design method–2007 Update with new Internet, visual, and mixed-mode guide* (John Wiley & Sons).

Fleiss JL (1971) Measuring nominal scale agreement among many raters. *Psychological Bulletin* 76(5):378.

Forcier J, Bissex P, Chun WJ (2008) *Python web development with Django* (Addison-Wesley Professional).

Goldberg LR (1970) Man versus model of man: A rationale, plus some evidence, for a method of improving on clinical inferences. *Psychological bulletin* 73(6):422.

Goldstein IM, Lawrence J, Miner AS (2017) Human-machine collaboration in cancer and beyond: The centaur care model. *JAMA Oncology* 3(10):1303–1304.

Goyal A, Acharya P, Pothuru S, Lahan S, Ranka S, Dalia T, Taduru S, Sauer AJ, Haglund N, Vidic A, et al. (2021) Thirty-day readmissions rates and causes among patients after heart transplant: Insights from the nationwide readmissions database. *Circulation* 144(Suppl_1):A13997–A13997.

Grand-Clément J, Pauphilet J (2024) The best decisions are not the best advice: Making adherence-aware recommendations. *Management Science* .

Harvey N, Fischer I (1997) Taking advice: Accepting help, improving judgment, and sharing responsibility. *Organizational Behavior and Human Decision Processes* 70(2):117–133.

Head T, Kumar M, Nahrstaedt H, Louppe G, Shcherbatyi I (2020) Scikit-optimize/scikit-optimize. *(version 0.8. 1)* .

Holzinger A (2016) Interactive machine learning for health informatics: when do we need the human-in-the-loop? *Brain Informatics* 3(2):119–131.

Ibrahim R, Kim SH, Tong J (2021) Eliciting human judgment for prediction algorithms. *Management Science* 67(4):2314–2325.

Imai K, Jiang Z, Greiner J, Halen R, Shin S (2020) Experimental evaluation of algorithm-assisted human decision-making: Application to pretrial public safety assessment. *arXiv preprint arXiv:2012.02845* .

Jalowiec A, Grady KL, White-Williams C (2008) Predictors of rehospitalization time during the first year after heart transplant. *Heart & Lung* 37(5):344–355.

Jencks SF, Williams MV, Coleman EA (2009) Rehospitalizations among patients in the medicare fee-for-service program. *New England Journal of Medicine* 360(14):1418–1428.

Jussupow E, Benbasat I, Heinzl A (2020) Why are we averse towards algorithms? a comprehensive literature review on algorithm aversion. *ECIS 2020 Proceedings* 168, URL `https://aisel.aisnet.org/ecis2020_rp/168`.

Karelaia N, Hogarth RM (2008) Determinants of linear judgment: A meta-analysis of lens model studies. *Psychological bulletin* 134(3):404.

Kasparov G (2010) The chess master and the computer. *The New York Review of Books* 57(2):16–19.

Kawaguchi K (2021) When will workers follow an algorithm? a field experiment with a retail business. *Management Science* 67(3):1670–1695.

Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D (2019) Key challenges for delivering clinical impact with artificial intelligence. *BMC medicine* 17:1–9.

King EA, Bowring MG, Massie AB, Kucirka LM, McAdams-DeMarco MA, Al-Ammary F, Desai NM, Segev DL (2017) Mortality and graft loss attributable to readmission following kidney transplantation: immediate and long-term risk. *Transplantation* 101(10):2520.

LeBlanc M, Tibshirani R (1996) Combining estimates in regression and classification. *Journal of the American Statistical Association* 91(436):1641–1650.

Logg JM, Minson JA, Moore DA (2019) Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes* 151:90–103.

Lovell MC (2008) A simple proof of the fwl theorem. *The Journal of Economic Education* 39(1):88–91.

Luan S, Singh S, Dai T (2024) Algorithmic bias and physician liability. *Available at SSRN 5046254* .

Lubetzky M, Yaffe H, Chen C, Ali H, Kayler LK (2016) Early readmission after kidney transplantation: examination of discharge-level factors. *Transplantation* 100(5):1079–1085.

Lundberg S, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, Katz R, Himmelfarb J, Bansal N, Lee SI (2020) From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence* 2(1):2522–5839.

Mann S, Naylor KL, McArthur E, Kim SJ, Knoll G, Zaltzman J, Treleaven D, Ouedraogo A, Jevnikar A, Garg AX (2020) Projecting the number of posttransplant clinic visits with a rise in the number of kidney transplants: a case study from ontario, canada. *Canadian Journal of Kidney Health and Disease* 7:2054358119898552.

McKinney SM, Sieniek M, Godbole V, Godwin J, Antropova N, Ashrafian H, Back T, Chesus M, Corrado GS, Darzi A, et al. (2020) International evaluation of an ai system for breast cancer screening. *Nature* 577(7788):89–94.

McLaughlin B, Spiess J (2022) Algorithmic assistance with recommendation-dependent preferences. *arXiv preprint arXiv:2208.07626* .

Munshi VN, Saghafian S, Cook CB, Werner KT, Chakkera HA (2020) Comparison of post-transplantation diabetes mellitus incidence and risk factors between kidney and liver transplantation patients. *PloS one* 15(1):e0226873.

Muthukrishnan R, Rohini R (2016) Lasso: A feature selection technique in predictive modeling for machine learning. *2016 IEEE international conference on advances in computer applications (ICACA)*, 18–20 (Ieee).

Orfanoudaki A, Cook CB, Saghafian S, Castro J, Kosiorek HE, Chakkera HA (2023) Diabetes mellitus and blood glucose variability increases the 30-day readmission rate after kidney transplantation. *Clinical Transplantation* e15177.

Panch T, Mattie H, Celi LA (2019) The "inconvenient truth" about ai in healthcare. *NPJ Digital Medicine* 2(1):1–3.

Patel MS, Mohebali J, Shah JA, Markmann JF, Vagefi PA (2016) Readmission following liver transplantation: an unwanted occurrence but an opportunity to act. *HPB* 18(11):936–942.

Patki N, Wedge R, Veeramachaneni K (2016) The synthetic data vault. *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, 399–410 (IEEE).

Peng K, Garg N, Kleinberg J (2024) A no free lunch theorem for human-ai collaboration. *arXiv preprint arXiv:2411.15230* .

Reardon S (2019) Rise of robot radiologists. *Nature* 576(7787):S54–S58.

Saghafian S (2024) Ambiguous dynamic treatment regimes: A reinforcement learning approach. *Management Science* 70(9):5667–5690.

Saghafian S, Hopp WJ (2019) The role of quality transparency in health care: Challenges and potential solutions. *NAM perspectives* 2019.

Saghafian S, Hopp WJ (2020) Can public reporting cure healthcare? The role of quality transparency in improving patient–provider alignment. *Operations Research* 68(1):71–92.

Saghafian S, Idan L (2024) Effective Generative AI: The Human-Algorithm Centaur. *Harvard Data Science Review* (Special Issue 5), https://hdsr.mitpress.mit.edu/pub/3rvlzjtw.

Sawyer J (1966) Measurement and prediction, clinical and statistical. *Psychological bulletin* 66(3):178.

Serrano OK, Vock DM, Chinnakotla S, Dunn TB, Kandaswamy R, Pruett TL, Feldman R, Matas AJ, Finger EB (2019) The relationships between cold ischemia time, kidney transplant length of stay, and transplant-related costs. *Transplantation* 103(2):401–411.

Sudhakar S, Zhang W, Kuo YF, Alghrouz M, Barbajelata A, Sharma G (2015) Validation of the readmission risk score in heart failure patients at a tertiary hospital. *Journal of Cardiac Failure* 21(11):885–891.

Sun J, Zhang DJ, Hu H, Van Mieghem JA (2022) Predicting human discretion to adjust algorithmic prescription: A large-scale field experiment in warehouse operations. *Management Science* 68(2):846–865.

Tian F, Liu D, Wei N, Fu Q, Sun L, Liu W, Sui X, Tian K, Nemeth G, Feng J, et al. (2024) Prediction of tumor origin in cancers of unknown primary origin with cytology-based deep learning. *Nature Medicine* 1–11.

Weiss AJ, Jiang HJ (2021) Overview of clinical conditions with frequent and costly hospital readmissions by payer, 2018: statistical brief# 278 .

Werner KT, Mackey PA, Castro JC, Carey EJ, Chakkera HA, Cook CB (2016) Hyperglycemia during the immediate period following liver transplantation. *Future Science OA* 2(1).

Wu X, Xiao L, Sun Y, Zhang J, Ma T, He L (2022) A survey of human-in-the-loop for machine learning. *Future Generation Computer Systems* .

Electronic Companion

EC.1. General Notation

Table EC.1 Summary of Notations

Notation	Description
N	Total number of observations
$\mathbf{x}_i \subseteq \mathbb{R}^d$	Feature vector for observation i with d dimensions
$y_i \subseteq \mathbb{R}$	Outcome variable for observation i
$f(\mathbf{x}_i)$	Baseline ML model predicting the outcome y with output $\hat{y}$
$l(\mathbf{x}_i)$	Prediction model for human experts' risk perception without ML predictions
$h(\mathbf{x}, \mathbf{z}, f(\mathbf{x}))$	Human risk perception model after observing ML predictions, incorporating expert heterogeneity
$q(\mathbf{x}_i, l(\mathbf{x}_i))$	Centaur model combining ML and human predictions to generate the final outcome prediction
$\mathbf{X} \in \mathbb{R}^{N \times d}$	Matrix of features for all N observations, where each row corresponds to $\mathbf{x}_i$
$\mathbf{y} \in \mathbb{R}^{N \times 1}$	Vector of outcomes for all N observations
$(\mathbf{X}_N, \mathbf{y}_N)$	Retrospective dataset containing all N observations
$(\mathbf{X}_T, \mathbf{y}_T)$	Training dataset derived from $(\mathbf{X}_N, \mathbf{y}_N)$
$(\mathbf{X}_E, \mathbf{y}_E)$	Testing dataset derived from $(\mathbf{X}_N, \mathbf{y}_N)$
$S \subset E$	Subset of testing data used in the online survey
$R = E \setminus S$	Remaining subset of testing data not included in the survey
$\mathbf{Z}_S$	Characteristics of human experts for observations in S
$\mathbf{Z}_R$	Simulated characteristics of human experts for observations in R
r_j	Human expert's prediction for observation j without observing ML predictions
w_j	Human expert's prediction for observation j after observing ML predictions
$\mathbf{r}_S$	Vector of human predictions (without ML) for observations in S
$\mathbf{w}_S$	Vector of human predictions (with ML) for observations in S
$\hat{\mathbf{r}} = l(\mathbf{X})$	Model trained on $\mathbf{r}_S$ to predict human risk perception without ML
$\hat{\mathbf{w}} = h(\mathbf{X}, \mathbf{Z}, f(\mathbf{X}))$	Model trained on $\mathbf{w}_S$ to predict human risk perception with ML and expert heterogeneity
$q(\mathbf{X}, l(\mathbf{X}))$	Centaur model integrating human risk perception with baseline ML predictions
$\hat{\mathbf{w}} = h(\mathbf{X}, \mathbf{Z}, q(\mathbf{X}, l(\mathbf{X})))$	Final centaur model used to evaluate performance in practice
$(\mathbf{X}_R, \mathbf{y}_R)$	Unseen portion of the testing data used to evaluate centaur model performance

EC.2. Centaur Model Proofs

This section provides the details of our technical results described in Section 3.

EC.2.1. Proof of Proposition 1

Proof of Proposition 1. We start by expanding the squared loss term for a convex combination of $f(\mathbf{X})$ and $l(\mathbf{X})$ to get

$$\begin{aligned}(y_i - (1-\alpha)f(\mathbf{x}_i) - \alpha l(\mathbf{x}_i))^2 &= (y_i - f(\mathbf{x}_i) + \alpha(f(\mathbf{x}_i) - l(\mathbf{x}_i)))^2 \\ &= (y_i - f(\mathbf{x}_i))^2 + 2\alpha(f(\mathbf{x}_i) - l(\mathbf{x}_i))(y_i - f(\mathbf{x}_i)) + \alpha^2 (f(\mathbf{x}_i) - l(\mathbf{x}_i))^2 .\end{aligned}$$

If we choose to average the predictions (i.e., $\alpha = \frac{1}{2}$), the squared loss becomes:

$$\begin{aligned}(y_i - v(\mathbf{x}_i))^2 &= \left(y_i - \frac{f(\mathbf{x}_i) + l(\mathbf{x}_i)}{2}\right)^2 \\ &= (y_i - f(\mathbf{x}_i))^2 + \frac{1}{2}(f(\mathbf{x}_i) - l(\mathbf{x}_i))(y_i - f(\mathbf{x}_i)) + \frac{1}{4}(f(\mathbf{x}_i) - l(\mathbf{x}_i))^2 .\end{aligned}$$

Rearranging, we can express the loss relative to $f(\mathbf{x}_i)$ as:

$$(y_i - f(\mathbf{x}_i))^2 = \left(y_i - \frac{f(\mathbf{x}_i)+l(\mathbf{x}_i)}{2}\right)^2 - \frac{1}{2}(f(\mathbf{x}_i)-l(\mathbf{x}_i))(y_i - f(\mathbf{x}_i)) - \frac{1}{4}\left(f(\mathbf{x}_i)-l(\mathbf{x}_i)\right)^2.$$

From Proposition EC.2, we substitute our expression into the minimization problem to yield

$$\begin{aligned}&\min_{\boldsymbol{\theta}} E\left[\frac{1}{N}\sum_{n=1}^{N}\left(y_i - q(\mathbf{x}_i, l(\mathbf{X}))\right)^2\right]\\ &= \mathsf{MSE}(\mathbf{y}, v(\mathbf{X})) - \frac{1}{N}\sum_{n=1}^{N}\left(E\big[(f(\mathbf{x}_i)-l(\mathbf{x}_i))(y_i - f(\mathbf{x}_i))\big] - \frac{1}{4}\left(f(\mathbf{x}_i)-l(\mathbf{x}_i)\right)^2\right) - \sigma^2_{l(\mathbf{X})}(\theta^*_{d+1})^2,\end{aligned}$$

where $\sigma^2_{l(\mathbf{X})}$ is the variance of $l(\mathbf{X})$ and θ^*_{d+1} is the least-squares solution for $q(\mathbf{x}, l(\mathbf{X}))$.

EC.2.2. Proof of Theorem 1

Proof of Theorem 1. We observe that

$$\min_{\boldsymbol{\beta}}\left\{\mathbb{E}\left[\sum_{i=1}^{N}\left(y_i - q\big(\mathbf{x}_i, l(\mathbf{x}_i)\big)\right)^2\right]\right\} \leq \min_{\alpha}\left\{\mathbb{E}\left[\sum_{i=1}^{N}\left(y_i - \alpha f(\mathbf{x}_i) - (1-\alpha)l(\mathbf{x}_i)\right)^2\right] : \sum_{i=1}^{2}\alpha_i = 1, \alpha \geq 0\right\}.$$

since we can always set $\sum_{m=1}^{M_1}\theta_{1,m} = \alpha_2\sum_{m=1}^{M_1}\beta_{1,m}, \sum_{m=1}^{M_2}\theta_{2,m} = \alpha_2\sum_{m=1}^{M_2}\beta_{2,m}, \ldots, \sum_{m=1}^{M_d}\theta_{dm} = \alpha_2\sum_{m=1}^{M_d}\theta_{dm}$ and $0 \leq \alpha_2 \leq 1$. By monotonicity of expectation, we obtain our desired result.

EC.2.3. Proof of Theorem 2

Proof of Theorem 2. Note that the mean squared error (MSE) for any predictor is

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[y_i^2] - \frac{2}{N}\sum_{i=1}^{N}\Big(\mathrm{Cov}(y_i, \hat{y}_i) + \mathbb{E}[y_i]\mathbb{E}[\hat{y}_i]\Big) + \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[\hat{y}_i^2],$$

where $\hat{y}_i$ is the prediction for y_i. By substitution, we write the MSE for $f(\mathbf{X})$, $l(\mathbf{X})$, and $q(\mathbf{x}, l(\mathbf{X}))$ explicitly. For $f(\mathbf{X})$, we have

$$\mathrm{MSE}(\mathbf{y}, f(\mathbf{X})) = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[y_i^2] - \frac{2}{N}\sum_{i=1}^{N}\Big(\mathrm{Cov}(y_i, f(\mathbf{x}_i)) + \mathbb{E}[y_i]\mathbb{E}[f(\mathbf{x}_i)]\Big) + \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[f(\mathbf{x}_i)^2].$$

Similarly, for $l(\mathbf{X})$ we have

$$\mathrm{MSE}(\mathbf{y}, l(\mathbf{X})) = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[y_i^2] - \frac{2}{N}\sum_{i=1}^{N}\Big(\mathrm{Cov}(y_i, l(\mathbf{x}_i)) + \mathbb{E}[y_i]\mathbb{E}[l(\mathbf{x}_i)]\Big) + \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[l(\mathbf{x}_i)^2].$$

Lastly, for the centaur predictor $q(\mathbf{x}, l(\mathbf{X}))$, we have

$$\mathrm{MSE}(\mathbf{y}, q(\mathbf{x}, l(\mathbf{X}))) = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[y_i^2] - \frac{2}{N}\sum_{i=1}^{N}\Big(\mathrm{Cov}(y_i, q(\mathbf{x}_i, l(\mathbf{x}_i))) + \mathbb{E}[y_i]\mathbb{E}[q(\mathbf{x}_i, l(\mathbf{x}_i))]\Big) + \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[q(\mathbf{x}_i, l(\mathbf{x}_i))^2].$$

Next, we observe that

$$\begin{aligned}\rho_{(l(\mathbf{X}),\mathbf{y})|\mathbf{x}} &= \frac{\mathrm{Cov}(\mathbf{y}, l(\mathbf{X})) - \mathrm{Cov}(\mathbf{y}, \hat{l}(\mathbf{X})) - \mathrm{Cov}(f(\mathbf{X}), l(\mathbf{X})) + \mathrm{Cov}(f(\mathbf{X}), \hat{l}(\mathbf{X}))}{\sigma_{M_{\mathbf{X}\cdot\mathbf{y}}}\sigma_{M_{\mathbf{X}}\cdot l(\mathbf{X})}}\\ &= \frac{\mathrm{MSE}(\mathbf{y}, \hat{l}(\mathbf{X})) - \mathrm{MSE}(\mathbf{y}, l(\mathbf{X})) + \mathrm{MSE}(f(\mathbf{X}), l(\mathbf{X})) - \mathrm{MSE}(f(\mathbf{X}), \hat{l}(\mathbf{X}))}{2\sigma_{M_{\mathbf{X}\cdot\mathbf{y}}}\sigma_{M_{\mathbf{X}}\cdot l(\mathbf{X})}}\end{aligned}$$

Therefore, we have

$$\begin{aligned}
&\mathrm{MSE}(f(\mathbf{X}), l(\mathbf{X})) - \mathrm{MSE}(\mathbf{y}, l(\mathbf{X})) \\
&= \frac{2}{N}\sum_{i=1}^{N} \mathbb{E}\Big[(f(\mathbf{x}_i) - \mathbf{y}_i)(\mathbf{y}_i + f(\mathbf{x}_i) - 2l(\mathbf{x}_i))\Big] - \mathrm{MSE}(f(\mathbf{X}), l(\mathbf{X})) + \mathrm{MSE}(\mathbf{y}, l(\mathbf{X})), \\
&\mathrm{MSE}(\mathbf{y}, \hat{l}(\mathbf{X})) - \mathrm{MSE}(f(\mathbf{X}), \hat{l}(\mathbf{X})) \\
&= \frac{2}{N}\sum_{i=1}^{N} \mathbb{E}\Big[(f(\mathbf{x}_i) - \mathbf{y}_i)(\hat{l}(\mathbf{x}_i) - l(\mathbf{x}_i) -)\Big] + \mathrm{MSE}(\mathbf{y}, l(\mathbf{X})) - \mathrm{MSE}(f(\mathbf{X}), l(\mathbf{X})).
\end{aligned}$$

Finally, substituting into the expressions for $\rho_{(l(\mathbf{X}),\mathbf{y})|\mathbf{x}}$ and simplifying, we find:

$$\rho_{(l(\mathbf{X}),\mathbf{y})|\mathbf{x}} = \frac{\mathrm{MSE}(\mathbf{y}, l(\mathbf{X})) - \mathrm{MSE}(f(\mathbf{X}), l(\mathbf{X})) + \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}\Big[(f(\mathbf{x}_i) - y_i)(y_i + \hat{l}(\mathbf{x}_i) + f(\mathbf{x}_i) - 3l(\mathbf{x}_i))\Big]}{\sigma_{M_{\mathbf{X}\cdot\mathbf{y}}}\sigma_{M_{\mathbf{X}}\cdot l(\mathbf{X})}}.$$

Lastly, by using Proposition EC.2, we obtain our desired result.

EC.2.4. Proof of Proposition 2

Proof of Proposition 2. **Case 1.** If $y_i - l(\mathbf{x}_i) = 0$ for all $i \in [N]$, then we know that $\mathbf{y} = l^*(\mathbf{x})$ and $\hat{l}(\mathbf{X}) = f(\mathbf{X})$. We obtain

$$\begin{aligned}
&\rho_{(l(\mathbf{X}),\mathbf{y})|\mathbf{x}} \\
&= \frac{\mathrm{MSE}(\mathbf{y}, l(\mathbf{X})) - \mathrm{MSE}(f(\mathbf{X}), l(\mathbf{X})) + \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}\left[\big(f(\mathbf{x}_i) - y_i\big)\big(y_i + \hat{l}(\mathbf{x}_i) + f(\mathbf{x}_i) - 3l(\mathbf{x}_i)\big)\right]}{\sigma_{M_{\mathbf{x}}\cdot\mathbf{y}}\sigma_{M_{\mathbf{x}}\cdot l^*(\mathbf{x})}} \\
&= \frac{\mathrm{MSE}(\mathbf{y}, f(\mathbf{X}))}{\sigma^2_{M_{\mathbf{x}}\cdot\mathbf{y}}} \\
&= \frac{\mathrm{MSE}(f(\mathbf{X}), \mathbf{y})}{\frac{1}{N}\sum_{i=1}^{N} \mathbb{E}\left[\big(y_i - f(\mathbf{x}_i) - \mathbb{E}[y_i - f(\mathbf{x}_i)]\big)^2\right]} \\
&= \frac{\mathrm{MSE}(\mathbf{y}, f(\mathbf{X}))}{\frac{1}{N}\sum_{i=1}^{N} \mathbb{E}\left[\big(y_i - f(\mathbf{x}_i)\big)^2\right]} \\
&= 1,
\end{aligned}$$

where the second-to-last line follows from $\mathbb{E}[f(\mathbf{x}_i)] = y_i$ by the definition of least squares. Moreover, we have

$$\frac{\sigma^2_{M_{\mathbf{x}}\cdot\mathbf{y}}}{\sigma^2_{M_{\mathbf{x}}\cdot l^*(\mathbf{x})}} = \frac{\sigma^2_{M_{\mathbf{x}}\cdot\mathbf{y}}}{\sigma^2_{M_{\mathbf{x}}\cdot\mathbf{y}}} = 1.$$

Thus, we have

$$\min_{\boldsymbol{\beta}} \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\big(y_i - q(\mathbf{x}_i, l(\mathbf{x}_i))\big)^2\right] = \min_{\boldsymbol{\beta}} \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\big(y_i - f(\mathbf{x}_i)\big)^2\right] - \sigma^2_{l(\mathbf{x}_i)}.$$

Case 2. If $f(\mathbf{x}_i) - l(\mathbf{x}_i) = 0$ then we know that $f(\mathbf{x}_i) = l(\mathbf{x}_i)$ and $f(\mathbf{x}_i) = \hat{l}(\mathbf{x}_i)$. We obtain

$$
\begin{aligned}
&\rho_{(l(\mathbf{X}),\mathbf{y})|\mathbf{x}} \\
&= \frac{\mathrm{MSE}(\mathbf{y}, l(\mathbf{X})) - \mathrm{MSE}(f(\mathbf{X}), l(\mathbf{X})) + \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}\left[\big(f(\mathbf{x}_i) - y_i\big)\big(y_i + \hat{l}(\mathbf{x}_i) + f(\mathbf{x}_i) - 3l(\mathbf{x}_i)\big)\right]}{\sigma_{M_{\mathbf{x}}\cdot\mathbf{y}}\sigma_{M_{\mathbf{x}}\cdot l^*(\mathbf{x})}} \\
&= \frac{\mathrm{MSE}(\mathbf{y}, l(\mathbf{X})) - \mathrm{MSE}(\mathbf{y}, l(\mathbf{X}))}{\sigma^2_{M_{\mathbf{x}}\cdot\mathbf{y}}} \\
&= \frac{\mathrm{MSE}(\mathbf{y}, l(\mathbf{X})) - \mathrm{MSE}(\mathbf{y}, f(\mathbf{X}))}{\sigma^2_{M_{\mathbf{x}}\cdot\mathbf{y}}} \\
&= 0,
\end{aligned}
$$

Thus, we have

$$
\min_{\boldsymbol{\beta}} \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\big(y_i - q(\mathbf{x}_i, l(\mathbf{x}_i))\big)^2\right] = \min_{\boldsymbol{\beta}} \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\big(y_i - f(\mathbf{x}_i)\big)^2\right].
$$

Case 3. If $y_i - l(\mathbf{x}_i) = C \in \mathcal{R}$ and $f(\mathbf{x}_i) - l(\mathbf{x}_i)) = C \in \mathcal{R}$ for all $i \in [N]$, then we note that

$$
\begin{aligned}
&\mathrm{MSE}(\mathbf{y}, l(\mathbf{X})) - \mathrm{MSE}(f(\mathbf{X}), l(\mathbf{X})) + \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}\left[\big(f(\mathbf{x}_i) - y_i\big)\big(y_i + \hat{l}^*(\mathbf{x}_i) + f(\mathbf{x}_i) - 3l(\mathbf{x}_i)\big)\right] \\
&= \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[\big(y_i - \hat{y}_i\big)^2\right] + \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[\big(f(\mathbf{x}_i) - y_i\big)\big(y_i - l(\mathbf{x}_i)\big)\right] \\
&\quad - \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[\big(f(\mathbf{x}_i) - l(\mathbf{x}_i)\big)^2\right] + \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[\big(f(\mathbf{x}_i) - y_i\big)\big(f(\mathbf{x}_i) - l(\mathbf{x}_i)\big)\right] \\
&\quad + \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[\big(f(\mathbf{x}_i) - y_i\big)\big(\hat{l}^*(\mathbf{x}_i) - l(\mathbf{x}_i)\big)\right] \\
&= \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[\big(f(\mathbf{x}_i) - y_i\big)\big(\hat{l}^*(\mathbf{x}_i) - l(\mathbf{x}_i)\big)\right].
\end{aligned}
$$

Let $l(\mathbf{x}_i) = \alpha y_i + \sqrt{1-\alpha^2}\epsilon_i$, where ϵ_i is independent of y_i and $\mathbf{x}_i$, and $\sigma^2_{\epsilon_i} = \sigma^2_{y_i}$, with $\alpha \in [-1, 1]$. Notice that $\rho_{y_i, l(\mathbf{x}_i)} = \alpha$. If $\hat{l}^*(\mathbf{x}_i) = \mathbf{x}_i\theta$, then by the method of least squares, we have $\theta^* = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T l(\mathbf{X})$, so

$$
\begin{aligned}
l(\mathbf{x}_i) - \hat{l}^*(\mathbf{x}_i) &= \alpha y_i + \sqrt{1-\alpha^2}\epsilon_i - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\big(\alpha y_i + \sqrt{1-\alpha^2}\epsilon_i\big) \\
&= \alpha y_i + \sqrt{1-\alpha^2}\epsilon_i - \alpha f(\mathbf{x}_i) - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\sqrt{1-\alpha^2}\epsilon_i \\
&= \alpha(y_i - f(\mathbf{x}_i)) + \sqrt{1-\alpha^2}\epsilon_i(I - H),
\end{aligned}
$$

where $H = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ is the projection matrix. Finally, we compute

$$\begin{aligned}
&\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[\big(y_i - f(\mathbf{x}_i)\big)\big(l(\mathbf{x}_i) - \hat{l}^*(\mathbf{x}_i)\big)\right] \\
&= \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[\big(y_i - f(\mathbf{x}_i)\big)\big(\alpha(y_i - f(\mathbf{x}_i)) + \sqrt{1-\alpha^2}\epsilon_i(I-H)\big)\right] \\
&= \alpha\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[\big(y_i - f(\mathbf{x}_i)\big)^2\right] + \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[\big(y_i - f(\mathbf{x}_i)\big)\sqrt{1-\alpha^2}\epsilon_i(I-H)\right] \\
&= \alpha\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[\big(y_i - f(\mathbf{x}_i)\big)^2\right] \\
&= \rho_{y_i,l(\mathbf{x}_i)}\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[\big(y_i - f(\mathbf{x}_i)\big)^2\right],
\end{aligned}$$

where the second line follows from ϵ_i being independent of y_i and $\mathbf{x}_i$, and the last line uses $\rho_{y_i,l(\mathbf{x}_i)} = \alpha$. Thus, we have

$$\begin{aligned}
\min_{\boldsymbol{\beta}}\mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\big(y_i - q(\mathbf{x}_i, l(\mathbf{x}_i))\big)^2\right] = \min_{\boldsymbol{\beta}}\mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\big(y_i - f(\mathbf{x}_i)\big)^2\right] \\
- \left(\rho_{Y,h(X)}\cdot\mathsf{MSE}\Big(\mathbf{y}, f(\mathbf{X})\Big)\cdot\frac{\sigma_{l(\mathbf{X})}}{\sigma^2_{M_{\mathbf{X}}\cdot l(\mathbf{X})}}\right)^2.
\end{aligned}$$

EC.2.5. Proposition EC.1

PROPOSITION EC.1. *Let* $\gamma \in \mathbb{R}^{[0,1]}$

$$\lim_{\gamma\to 0}\Delta = \begin{cases} 0 & |\rho_{\mathbf{y},l(\mathbf{X})}| = 1-\gamma \text{ and } |\rho_{l(\mathbf{X}),f(\mathbf{X})}| = 1-\gamma \\ 0 & |\rho_{\mathbf{y},l(\mathbf{X})}| = \gamma \text{ and } |\rho_{l(\mathbf{X}),f(\mathbf{X})}| = 1-\gamma \\ C_1 & |\rho_{\mathbf{y},l(\mathbf{X})}| = 1-\gamma \text{ and } |\rho_{l(\mathbf{X}),f(\mathbf{X})}| = \gamma \\ C_2 & |\rho_{\mathbf{y},l(\mathbf{X})}| = \gamma \text{ and } |\rho_{l(\mathbf{X}),f(\mathbf{X})}| = \gamma \end{cases}$$

where $C_1, C_2 \in \mathbb{R}^+$ *are some constants.*

Proof of Proposition EC.1. **Case 1.** Assuming $l(\mathbf{X})$ is strongly correlated with both $\mathbf{y}$ and $f(\mathbf{X})$, as $\epsilon \to 0$, we get

$$\begin{aligned}
\rho_{(l(\mathbf{X}),\mathbf{y})|\mathbf{x}} &= \frac{\rho_{\mathbf{y},l(\mathbf{X})}\sigma_{\mathbf{y}}\sigma_{l(\mathbf{X})} - \rho_{\mathbf{y},\hat{l}(\mathbf{X})}\sigma_{\mathbf{y}}\sigma_{\hat{l}(\mathbf{X})} - \rho_{f(\mathbf{X}),l(\mathbf{X})}\sigma_{f(\mathbf{X})}\sigma_{l(\mathbf{X})} + \rho_{f(\mathbf{X}),\hat{l}(\mathbf{X})}\sigma_{f(\mathbf{X})}\sigma_{\hat{l}(\mathbf{X})}}{\sigma_{M_{\mathbf{X}}\cdot\mathbf{y}}\sigma_{M_{\mathbf{X}}\cdot l(\mathbf{X})}} \\
&= \frac{\sigma_{\mathbf{y}}\sigma_{l(\mathbf{X})} - \sigma_{\mathbf{y}}\sigma_{\hat{l}(\mathbf{X})} - \sigma_{f(\mathbf{X})}\sigma_{l(\mathbf{X})} + \sigma_{f(\mathbf{X})}\sigma_{\hat{l}(\mathbf{X})}}{\sigma_{M_{\mathbf{X}}\cdot\mathbf{y}}\sigma_{M_{\mathbf{X}}\cdot l(\mathbf{X})}} \\
&= \frac{\big(\sigma_{\mathbf{y}} - \sigma_{f(\mathbf{X})}\big)\Big(\sigma_{l(\mathbf{X})} - \sigma_{\hat{l}(\mathbf{X})}\Big)}{\sigma_{M_{\mathbf{X}}\cdot\mathbf{y}}\sigma_{M_{\mathbf{X}}\cdot l(\mathbf{X})}} \\
&= 0.
\end{aligned}$$

Here, we use the assumption that perfect correlation implies linearity of all three models. Thus, we have

$$\lim_{\epsilon\to 0}\left\{\mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\Big(\mathbf{y}_i - q(\mathbf{x}_i, l(\mathbf{x}_i))\Big)^2\right]\right\} = \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\Big(\mathbf{y}_i - f(\mathbf{x}_i)\Big)^2\right].$$

Case 2. Assuming $l(\mathbf{X})$ is weakly correlated with $\mathbf{y}$ and strongly correlated with $f(\mathbf{X})$, as $\epsilon \to 0$, we get

$$\begin{aligned}
\rho_{(l(\mathbf{X}),\mathbf{y})|\mathbf{x}} &= \frac{\rho_{\mathbf{y},l(\mathbf{X})}\sigma_{\mathbf{y}}\sigma_{l(\mathbf{X})} - \rho_{\mathbf{y},\hat{l}(\mathbf{X})}\sigma_{\mathbf{y}}\sigma_{\hat{l}(\mathbf{X})} - \rho_{f(\mathbf{X}),l(\mathbf{X})}\sigma_{f(\mathbf{X})}\sigma_{l(\mathbf{X})} + \rho_{f(\mathbf{X}),\hat{l}(\mathbf{X})}\sigma_{f(\mathbf{X})}\sigma_{\hat{l}(\mathbf{X})}}{\sigma_{M_{\mathbf{X}\cdot\mathbf{y}}}\sigma_{M_{\mathbf{X}}\cdot l(\mathbf{X})}} \\
&= \frac{-\sigma_{f(\mathbf{X})}\sigma_{l(\mathbf{X})} + \sigma_{f(\mathbf{X})}\sigma_{\hat{l}(\mathbf{X})}}{\sigma_{M_{\mathbf{X}\cdot\mathbf{y}}}\sigma_{M_{\mathbf{X}}\cdot l(\mathbf{X})}} \\
&= \frac{\sigma_{f(\mathbf{X})}\Big(\sigma_{\hat{l}(\mathbf{X})} - \sigma_{l(\mathbf{X})}\Big)}{\sigma_{M_{\mathbf{X}\cdot\mathbf{y}}}\sigma_{M_{\mathbf{X}}\cdot l(\mathbf{X})}} \\
&= 0.
\end{aligned}$$

The last line follows from $f(\mathbf{X}) = l(\mathbf{X})$ as $\epsilon \to 0$. Thus, we have

$$\lim_{\epsilon\to 0}\left\{\mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\Big(\mathbf{y}_i - q(\mathbf{x}_i, l(\mathbf{x}_i))\Big)^2\right]\right\} = \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\Big(\mathbf{y}_i - f(\mathbf{x}_i)\Big)^2\right].$$

Case 2c

In this case, we assume $l(\mathbf{X})$ is strongly correlated with $\mathbf{y}$ and weakly correlated with $f(\mathbf{X})$. As $\epsilon \to 0$, we observe that

$$\begin{aligned}
\rho_{(l(\mathbf{X}),\mathbf{y})|\mathbf{x}} &= \frac{\rho_{\mathbf{y},l(\mathbf{X})}\sigma_{\mathbf{y}}\sigma_{l(\mathbf{X})} - \rho_{\mathbf{y},\hat{l}(\mathbf{X})}\sigma_{\mathbf{y}}\sigma_{\hat{l}(\mathbf{X})} - \rho_{f(\mathbf{X}),l(\mathbf{X})}\sigma_{f(\mathbf{X})}\sigma_{l(\mathbf{X})} + \rho_{f(\mathbf{X}),\hat{l}(\mathbf{X})}\sigma_{f(\mathbf{X})}\sigma_{\hat{l}(\mathbf{X})}}{\sigma_{M_{\mathbf{X}\cdot\mathbf{y}}}\sigma_{M_{\mathbf{X}}\cdot l(\mathbf{X})}} \\
&= \frac{\sigma_{\mathbf{y}}\sigma_{l(\mathbf{X})} + \rho_{f(\mathbf{X}),\hat{l}(\mathbf{X})}\sigma_{f(\mathbf{X})}\sigma_{\hat{l}(\mathbf{X})}}{\sigma_{M_{\mathbf{X}\cdot\mathbf{y}}}\sigma_{M_{\mathbf{X}}\cdot l(\mathbf{X})}}.
\end{aligned}$$

The last line is due to the observation that if $l(\mathbf{X})$ and $f(\mathbf{X})$ are not correlated, then $\hat{l}(\mathbf{X})$ must also not be strongly correlated with $\mathbf{y}$. To see this, assume that $\hat{l}(\mathbf{X})$ is strongly correlated with $\mathbf{y}$ and that $l(\mathbf{X})$ and $f(\mathbf{X})$ are not correlated with $\mathbf{y}$. Since $\hat{l}(\mathbf{X})$ is a linear function of $\mathbf{x}$ and so is $f(\mathbf{X})$, then $f(\mathbf{X})$ would also need to be strongly correlated with $\mathbf{y}$, which is a contradiction. Thus, we have

$$\begin{aligned}
\min_{\boldsymbol{\beta}}\mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\Big(\mathbf{y}_i - q(\mathbf{x}_i, l(\mathbf{x}_i))\Big)^2\right] &= \min_{\boldsymbol{\beta}}\mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\Big(\mathbf{y}_i - f(\mathbf{x}_i)\Big)^2\right] - \sigma^2_{l(\mathbf{X})}\rho^2_{(l(\mathbf{X}),\mathbf{y})|\mathbf{x}}\frac{\sigma^2_{M_{\mathbf{X}\cdot\mathbf{y}}}}{\sigma^2_{M_{\mathbf{X}}\cdot l(\mathbf{X})}} \\
&= \min_{\boldsymbol{\beta}}\mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\Big(\mathbf{y}_i - f(\mathbf{x}_i)\Big)^2\right] - C_1.
\end{aligned}$$

Case 2d

In this case, we assume $l(\mathbf{X})$ is weakly correlated with $\mathbf{y}$ and $f(\mathbf{X})$. As $\epsilon \to 0$, we observe that

$$\begin{aligned}\rho_{(l(\mathbf{X}),\mathbf{y})|\mathbf{x}} &= \frac{\rho_{\mathbf{y},l(\mathbf{X})}\sigma_{\mathbf{y}}\sigma_{l(\mathbf{X})} - \rho_{\mathbf{y},\hat{l}(\mathbf{X})}\sigma_{\mathbf{y}}\sigma_{\hat{l}(\mathbf{X})} - \rho_{f(\mathbf{X}),l(\mathbf{X})}\sigma_{f(\mathbf{X})}\sigma_{l(\mathbf{X})} + \rho_{f(\mathbf{X}),\hat{l}(\mathbf{X})}\sigma_{f(\mathbf{X})}\sigma_{\hat{l}(\mathbf{X})}}{\sigma_{M_{\mathbf{X}}\cdot\mathbf{y}}\sigma_{M_{\mathbf{X}}\cdot l(\mathbf{X})}} \\ &= \frac{-\rho_{\mathbf{y},\hat{l}(\mathbf{X})}\sigma_{\mathbf{y}}\sigma_{\hat{l}(\mathbf{X})} + \rho_{f(\mathbf{X}),\hat{l}(\mathbf{X})}\sigma_{f(\mathbf{X})}\sigma_{\hat{l}(\mathbf{X})}}{\sigma_{M_{\mathbf{X}}\cdot\mathbf{y}}\sigma_{M_{\mathbf{X}}\cdot l(\mathbf{X})}} \\ &= \frac{\sigma_{\hat{l}(\mathbf{X})}\left(\rho_{f(\mathbf{X}),\hat{l}(\mathbf{X})}\sigma_{f(\mathbf{X})} - \rho_{\mathbf{y},\hat{l}(\mathbf{X})}\sigma_{\mathbf{y}}\right)}{\sigma_{M_{\mathbf{X}}\cdot\mathbf{y}}\sigma_{M_{\mathbf{X}}\cdot l(\mathbf{X})}}.\end{aligned}$$

Thus, we have

$$\begin{aligned}\min_{\boldsymbol{\beta}} \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\left(\mathbf{y}_i - q(\mathbf{x}_i, l(\mathbf{x}_i))\right)^2\right] &= \min_{\boldsymbol{\beta}} \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\left(\mathbf{y}_i - f(\mathbf{x}_i)\right)^2\right] - \sigma^2_{l(\mathbf{X})}\rho^2_{(l(\mathbf{X}),\mathbf{y})|\mathbf{x}}\frac{\sigma^2_{M_{\mathbf{X}}\cdot\mathbf{y}}}{\sigma^2_{M_{\mathbf{X}}\cdot l(\mathbf{X})}} \\ &= \min_{\boldsymbol{\beta}} \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\left(\mathbf{y}_i - f(\mathbf{x}_i)\right)^2\right] - C_2.\end{aligned}$$

EC.2.6. Proof of Lemma EC.1

LEMMA EC.1 **(Three-way Interaction)**. *Let Δ is the difference between minimal MSE of $f(\mathbf{X})$ and $q(\mathbf{X}, l(\mathbf{X}))$ with respect to y. We have then that*

$$\Delta = \left(\left(\rho_{\mathbf{y},l(\mathbf{X})}\sigma_{\mathbf{y}}\sigma_{l(\mathbf{X})} - \rho_{\mathbf{y},\hat{l}(\mathbf{X})}\sigma_{\mathbf{y}}\sigma_{\hat{l}(\mathbf{X})} - \rho_{f(\mathbf{X}),l(\mathbf{X})}\sigma_{f(\mathbf{X})}\sigma_{l(\mathbf{X})} + \rho_{f(\mathbf{X}),\hat{l}(\mathbf{X})}\sigma_{f(\mathbf{X})}\sigma_{\hat{l}(\mathbf{X})}\right)\cdot\frac{\sigma_{l(\mathbf{X})}}{\sigma^2_{M_{\mathbf{X}}\cdot l(\mathbf{X})}}\right)^2,$$

where $\hat{l}(\mathbf{X})$ is a linear model estimate of $l(\mathbf{X})$.

Proof of Lemma EC.1. Note that $M_{\mathbf{X}}\cdot\mathbf{y} = \mathbf{y} - f(\mathbf{X})$ and $M_{\mathbf{X}}\cdot l(\mathbf{X}) = l(\mathbf{X}) - \hat{l}(\mathbf{X})$, which implies

$$\rho_{l(\mathbf{X}),\mathbf{y}|\mathbf{x}} = \frac{\mathrm{Cov}(\mathbf{y} - f(\mathbf{X}), l(\mathbf{X}) - \hat{l}(\mathbf{X}))}{\sigma_{M_{\mathbf{X}}\cdot\mathbf{y}}\sigma_{M_{\mathbf{X}}\cdot l(\mathbf{X})}}.$$

We can further expand this to get

$$\begin{aligned}\rho_{l(\mathbf{X}),\mathbf{y}|\mathbf{x}} &= \frac{\mathrm{Cov}(\mathbf{y}, l(\mathbf{X})) - \mathrm{Cov}(\mathbf{y}, \hat{l}(\mathbf{X})) - \mathrm{Cov}(f(\mathbf{X}), l(\mathbf{X})) + \mathrm{Cov}(f(\mathbf{X}), \hat{l}(\mathbf{X}))}{\sigma_{M_{\mathbf{X}}\cdot\mathbf{y}}\sigma_{M_{\mathbf{X}}\cdot l(\mathbf{X})}} \\ &= \frac{\rho_{\mathbf{y},l(\mathbf{X})}\sigma_{\mathbf{y}}\sigma_{l(\mathbf{X})} - \rho_{\mathbf{y},\hat{l}(\mathbf{X})}\sigma_{\mathbf{y}}\sigma_{\hat{l}(\mathbf{X})} - \rho_{f(\mathbf{X}),l(\mathbf{X})}\sigma_{f(\mathbf{X})}\sigma_{l(\mathbf{X})} + \rho_{f(\mathbf{X}),\hat{l}(\mathbf{X})}\sigma_{f(\mathbf{X})}\sigma_{\hat{l}(\mathbf{X})}}{\sigma_{M_{\mathbf{X}}\cdot\mathbf{y}}\sigma_{M_{\mathbf{X}}\cdot l(\mathbf{X})}}.\end{aligned}$$

By Proposition EC.2, we can substitute our above derivation of the partial correlation to complete our argument.

EC.2.7. Additional Lemmas and Propositions

PROPOSITION EC.2. *If $q(\mathbf{x}, l(\mathbf{X}))$ is the centaur model and $f(\mathbf{X})$ is the ML model, then we have*

$$\min_{\boldsymbol{\theta}} \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\left(y_i - q(\mathbf{x}_i, l(\mathbf{x}_i))\right)^2\right] = \min_{\boldsymbol{\beta}} \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\left(y_i - f(\mathbf{x}_i)\right)^2\right] - \sigma^2_{l(\mathbf{X})}\left(\rho_{y,l(\mathbf{X})|\mathbf{x}}\frac{\sigma_{M_{\mathbf{X}}\cdot y}}{\sigma_{M_{\mathbf{X}}\cdot l(\mathbf{X})}}\right)^2,$$

where $\sigma^2_{l(\mathbf{X})}$ *is the variance of* $l(\mathbf{X})$ *and* $\theta^*_{d+1,1}$ *is from the least-square solution of* $q(\mathbf{x}, l(\mathbf{X}))$*. The term* $\rho_{y,l(\mathbf{X})|\mathbf{x}}$ *is the partial correlation between* y *and* $l(\mathbf{X})$ *after controlling for* $\mathbf{x}$*. The term* $\sigma_{M_{\mathbf{X}}\cdot y}$ *and* $\sigma_{M_{\mathbf{X}}\cdot l(\mathbf{X})}$ *are the standard deviation of residuals for* $M_{\mathbf{X}} \cdot l(\mathbf{X}) = l(\mathbf{X}) - f'(\mathbf{x})$ *and* $M_{\mathbf{X}} \cdot y = y - f(\mathbf{X})$*, respectively. Note that* $f'(\mathbf{x})$ *is a linear model estimate of* $l(\mathbf{X})$*.*

Proof of Proposition EC.2. Let $\theta^*_{d+1,1}$ be the least-square solution obtained for the model $q(\mathbf{x}, l(\mathbf{X}))$. Assume that $l(\mathbf{X})$ is standardized to have zero mean. We have then

$$
\begin{aligned}
&\min_{\boldsymbol{\beta}} \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\left(y_i - f(\mathbf{x}_i)\right)^2\right] \\
&= \min_{\boldsymbol{\beta}} \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\left(y_i - l(\mathbf{x}_i)\theta^*_{d+1,1} - \sum_{j=1}^{d}\sum_{m=1}^{M_j}\beta_{jm}\phi_m(\mathbf{x}_i) + l(\mathbf{x}_i)\theta^*_{d+1,1}\right)^2\right] \\
&= \min_{\boldsymbol{\theta}} \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\left(y_i - l(\mathbf{x}_i)\theta^*_{d+1,1} - \sum_{j=1}^{d}\sum_{m=1}^{M_j}\theta_{jm}\phi_m(\mathbf{x}_i) + l(\mathbf{x}_i)\theta^*_{d+1,1}\right)^2\right] \\
&= \min_{\boldsymbol{\theta}} \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\left(y_i - q(\mathbf{x}_i, l(\mathbf{x}_i)) + l(\mathbf{x}_i)\theta^*_{d+1,1}\right)^2\right] \\
&= \min_{\boldsymbol{\theta}} \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\left(y_i - q(\mathbf{x}_i, l(\mathbf{x}_i))\right)^2 + l(\mathbf{x}_i)\theta^*_{d+1,1}\Big(l(\mathbf{x}_i)\theta^*_{d+1,1} - 2\big(y_i - q(\mathbf{x}_i, l(\mathbf{x}_i))\big)\Big)\right] \\
&= \min_{\boldsymbol{\theta}} \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\left(y_i - q(\mathbf{x}_i, l(\mathbf{x}_i))\right)^2\right] + \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\left(l(\mathbf{x}_i)\theta^*_{d+1,1}\right)^2\right] \\
&= \min_{\boldsymbol{\theta}} \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\left(y_i - q(\mathbf{x}_i, l(\mathbf{x}_i))\right)^2\right] + \sigma^2_{l(\mathbf{X})}(\theta^*_{d+1,1})^2.
\end{aligned}
$$

The second-to-last line follows from the fact that the expected residuals are zero for the least-square solution. Lastly, we use the the Frisch-Waugh-Lovell theorem to obtain our desired result (Lovell 2008). Specifically, we know that θ^*_{d+1} is from the least-square solution of $\mathsf{MSE}(\mathbf{y}, q(\mathbf{X}, l(\mathbf{X})))$, then

$$\theta^*_{d+1} = \rho_{y,l(\mathbf{X})|\mathbf{x}}\frac{\sigma_{M_{\mathbf{X}}\cdot y}}{\sigma_{M_{\mathbf{X}}\cdot l(\mathbf{X})}}.$$

By substitution, we obtain our desired result.

EC.3. Extension of the Centaur Framework

We have as a continuation from Step 5 of our proposed Centaur Framework detailed in Section 4:

6. **Simulate samples of human experts:** We let $R = E \setminus S$ represent the set of observations that are part of the testing set but were not included in the survey. We use the characteristics of the survey respondents to generate a representative sample of synthetic experts $\mathbf{Z}_R$ (see Section 5.8). This simulated sample serves as a representative pool of the human decision-makers that work in the organization where the centaur model will be deployed. Alternatively,

if the survey collects responses for all observations in E, randomly partition the testing set in the subsets R and S to derive two independent populations for the centaur and human experts model derivation and validation, respectively.

7. **Evaluate the expected performance of the model for implementation in practice:** As the last step, we consider the model $\hat{\mathbf{w}} = h(\mathbf{X}, \mathbf{Z}, q(\mathbf{X}, l(\mathbf{X}))$, evaluated on the unseen portion of the testing set $(\mathbf{X}_R, \mathbf{y}_R)$ to measure the expected aggregate performance of the centaur model when used by human decision makers $q(\mathbf{X}, l(\mathbf{X}))$ (see Sections 5.8-5.9).

EC.4. Data Description

Variable	Distribution Information	Organ	Variable	Distribution Information	Organ
Outcome 30-Day Readmission	353.0 (23.0%)	All	**Organ Type**		
Recipient Information			Organ Kidney	1037.0 (67.5%)	All
Age	56.0 (45.0-64.0)	All	Organ Liver	364.0 (23.7%)	All
Gender Male	947.0 (61.6%)	All	Organ Heart	136 (8.85%)	All
Race White	1111.0 (72.3%)	All	**Recipient Insulin Treatment**		
Race Asian	83.0 (5.4%)	All	Basal and Bolus First 24hrs	150.0 (9.8%)	All
Race Black or African American	128.0 (8.3%)	All	Bolus First 24hrs	606.0 (39.4%)	All
Race Other	122.0 (7.9%)	All	None First 24hrs	772.0 (50.2%)	All
Not Hispanic or Latino	1181.0 (76.8%)	All	Basal and Bolus Middle 24hrs	406.0 (26.4%)	All
Body Mass Index	27.8 (24.2-31.9)	All	Bolus Middle 24hrs	697.0 (45.3%)	All
MSDRG Weight	3.3 (3.3-10.3)	All	None Middle 24hrs	429.0 (27.9%)	All
Length of Stay at Index Admission	4.0 (3.0-7.0)	All	Basal and Bolus Last 24hrs	260.0 (16.9%)	All
Donor Information			Bolus Last 24hrs	402.0 (26.2%)	All
Age	40.0 (27.0-53.0)	All	None Last 24hrs	861.0 (56.0%)	All
Gender Male	888.0 (57.8%)	All	IV Therapy	808.0 (52.6%)	All
Race White	1004.0 (65.3%)	All	**Transplantation Information**		
Race Asian	49.0 (3.2%)	All	Creatinine Value at Discharge	2.2 (1.1-4.9)	All
Race Black or African American	126.0 (8.2%)	All	DCD Controlled Donor	308.0 (44.0%)	Kidney, Liver
Race Hispanic or Latino	312.0 (20.3%)	All	EPTS at Transplant	0.4 (0.2-0.7)	Kidney
Race Other	45.0 (2.9%)	All	HLA Mismatch Level	4.0 (3.0-5.0)	All
Donor Deceased	1321.0 (86.0%)	All	Time on Dialysis prior to Transplant	992.0 (465.5-1729.5)	Kidney
Body Mass Index	27.1 (23.3-32.1)	All	Cold Ischemic Time (Hours)	17.9 (6.9-23.7)	Kidney
Recipient Metabolic Factors			Presence of Delayed Graft Function	489.0 (31.8%)	Kidney
History of Diabetes mellitus	595.0 (38.7%)	All	A Locus Mismatch Level	2.0 (1.0-2.0)	Liver, Heart
Average HbA1c Value	5.7 (5.1-6.9)	All	B Locus Mismatch Level	2.0 (1.0-2.0)	Liver, Heart
Hyperglycemia	1007.0 (65.5%)	All	DR Locus Mismatch Level	2.0 (1.0-2.0)	Liver, Heart
Hypoglycemia	260.0 (16.9%)	All	Graft Status Functioning	331.0 (21.5%)	Liver
% of BG Measurements above 180	13.9 (1.2-33.3)	All	Use of Inotropes prior to Transplant	69.0 (4.5%)	Heart
% of BG Measurements below 70	0.9 (0.0-1.3)	All	Functional Status at Listing	70.0 (50.0-80.0)	Liver, Heart
BG Average Value First 24hrs	145.0 (126.8-167.2)	All	Functional Status at Transplant	70.0 (40.0-80.0)	Liver, Heart
BG Average Value Middle 24hrs	146.0 (126.0-170.0)	All	MELD Score	18.0 (12.0-25.0)	Liver
BG Average Value Last 24hrs	143.0 (126.0-170.0)	All	Donation after Circulatory Death	105.0 (6.8%)	All
BG Maximum Value First 24hrs	190.5 (155.0-236.0)	All	LVAD Presence	50.0 (3.3%)	Heart
BG Maximum Value Middle 24hrs	173.0 (146.0-221.0)	All	Portal Vein Tumor Thrombus	74.0 (4.8%)	Liver
BG Maximum Value Last 24hrs	173.0 (149.0-221.2)	All	Wait List Status Code at Listing	12.0 (2.0-18.0)	Liver, Heart
BG Minimum Value First 24hrs	103.0 (85.0-125.0	All	Bilirubin at transplant	0.6 (0.4-0.9)	Heart
BG Minimum Value Middle 24hrs	119.0 (102.0-137.0)	All	Diagnosis Alcoholic Cirrhosis	64.0 (4.2%)	Liver
BG Minimum Value Last 24hrs	115.0 (99.0-134.0	All	Diagnosis Dilated Myopathy	98.0 (6.4%)	Liver
Range of BG Values First 24 hrs	84.0 (41.0-136.0)	All	Diagnosis Other Cirrhosis	88.0 (5.7%)	Liver
Range of BG Values Middle 24 hrs	52.0 (32.0-87.0)	All	Diagnosis Hepatoma and Cirrhosis	94.0 (6.1%)	Liver
Range of BG Values Last 24 hrs	58.0 (36.0-90.0)	All	Diagnosis Other	118.0 (7.7%)	Liver

Notes. For continuous variables, we report the average and the 95% confidence interval. In the case of binary variables, the table shows the count of observations where the feature is present and, in parentheses, the percentage over the entire population. The last column indicates for which organ(s) the variable is present. We define the following acronyms: BG: Blood Glucose (fasting plasma glucose levels); EPTS: Estimated Post Transplant Survival score; BMI: Body Mass Index; HbA1c: Hemoglobin A1c; HLA: Human Leukocyte Antigens; LVAD: Left Ventricular Assist Device; MSDRG: Medicare Severity-Diagnosis Related Group, MELD: Model for End-Stage Liver Disease.

Table EC.2 Summary statistics of all clinical features for the patient population.

EC.5. Machine Learning Model Comparison

Algorithm	Average AUC	95% CI
Regularized Regression	75.4%	(75.0%, 75.8%)
CART	78.3%	(85.9%, 80.7%)
Random Forests	82.7%	(81.2%, 84.2%)
XGBoost	84.0%	(83.0%, 85.0%)
SVM	78.4%	(75.0%, 78.8%)
MLP	79.2%	(76.2%, 82.2%)

Table EC.3 Summary of AUC performance of ML algorithms considered on the testing set.

EC.6. Survey Details

Figure EC.1 Illustration of the survey tool interface for an example liver patient.

EC.6.1. Survey Platform

The online survey was hosted on a secure and encrypted server. The study was also approved by the Mayo Clinic's Institutional Review Board. Participants first reviewed the study setting, including information regarding the patient population, the survey objective, and the quality of the ML model. Specifically, the study's landing page highlighted the proposed ML model's out-of-sample accuracy. On the same page, users were provided with instructions on how to submit their answers. To ensure a common interpretation of patient features, detailed definitions for each variable were made available. Subjects were randomly assigned to patients subject to the constraint that each patient could only be reviewed by the same expert at most once. Endocrinologists were assigned to all types of organs. However, transplantation experts were only assigned to patients that had received an organ of their specialty.

EC.6.2. Participants

In total, 38 experts submitted their responses to the survey. 68.42% were Doctors of Medicine (MDs) and 31.58% Advanced Practice Providers (APPs). We invited to the survey platform participants from the two primary clinical divisions (transplantation and endocrinology) that are responsible for patient care during solid organ transplantation. Across all participants, 31.58% of the experts specialized in transplantation while 68.42% were based at the endocrinology department. We measure the degree of professional experience as the time since the expert passed the board certification exam. The mean number of years of experience among the survey respondents was 17.26 with a standard deviation of 10.94. To complete the survey, each expert was shown five distinct patient cases randomized from the testing set of the sample population. Some providers chose to respond to fewer cases. Thus, the average number of patient records reviewed per expert was 3.47.

EC.6.3. User Interface

Given the high levels of workload and stress that medical practitioners face, we placed a lot of emphasis on the design of the user interface. We aimed to provide an intuitive platform to minimize the time needed to submit an informed response. An example is shown in Figure EC.1. As seen from this figure, we included a dashboard to illustrate BG measurements throughout the hospital stay and separate tables to summarize different types of patient and organ information. We programmed the user interface using the Django library in Python (Forcier et al. 2008).

EC.7. Human Expert Risk Perception: ML Models Comparison

In Table EC.4, we report the average MAE and Brier score in the testing set across for five bootstrapped partitions of the data.

Algorithm	MAE	Brier Score
Linear Regression	0.111	0.020
CART	0.216	0.070
Random Forests	0.148	0.034
XGBoost	0.191	0.048
SVM	0.127	0.029

Table EC.4 Summary of predictive performance of ML algorithms considered on the testing set for the human risk perception model.

EC.8. Human Risk Models Details

Independent Variable	**Regression Coefficient**	p**-value**
BG Minimum Value Middle 24hrs	-0.0043	<0.001
BG Average Value First 24hrs	0.0029	<0.001
% of BG Measurements above 180	0.0029	<0.001
Donor Age	0.0029	<0.005
Recipient Age	0.0027	<0.005
BG Maximum Value Last 24hrs	0.0022	<0.01
BG Maximum Value Middle 24hrs	0.0022	<0.01
Donor Body Mass Index	0.0015	<0.01
BG Minimum Value First 24hrs	-0.0015	<0.01
Recipient Body Mass Index	0.0015	<0.01

Table EC.5 Output summary of the regularized regression model. We report the resulting coefficients only for the reduced model with statistically significant t-tests values.

Independent Variable	**Regression Coefficient**	p**-value**	**2.5% Q**	**97.5% Q**
Constant	-0.5012	0.002	-0.808	-0.194
Recipient Age at Admission	0.0023	0.043	7.11E-05	0.0050
Recipient BMI	0.0081	0.006	0.0025	0.014
Creatinine Value at Discharge	0.0033	0.039	0.0018	0.0048
Average BG Value in Last 24 hrs	0.0015	0.001	0.001	0.0020
History of diabetes 2 mellitus	0.0161	0.0073	0.0076	0.0246
HbA1c at admission	0.0246	0.0101	0.015	0.0342

Table EC.6 Output summary of the updated linear regression model. We report the resulting coefficients only for the reduced model with statistically significant t-tests values.

EC.9. Reasoning and Participant Consensus

Patient Characteristic	**Fleiss κ**	***p*-value**
HLA mismatch level	0.654	$p \leq 0.05$
Donor BMI	0.304	$p \leq 0.05$
Presence of delayed graft function	0.264	$p \leq 0.01$
Donor age	0.224	$p \leq 0.01$
Cold Ischemic Time (Hours)	0.211	$p \leq 0.01$
Max BG Value	0.113	$p \leq 0.05$
Creatinine value at discharge	0.111	$p \leq 0.05$
History of diabetes 2 mellitus	0.094	$p \leq 0.01$
Organ	0.082	$p \leq 0.01$
HbA1c at admission	0.075	$p \leq 0.05$
Recipient race	0.069	$p \leq 0.05$
Recipient Age	0.062	$p \leq 0.05$
Presence of hyperglycemia during admission	0.031	$p \leq 0.05$

Table EC.7 Fleiss' κ measure of inter-rater agreement between human experts on patient clinical characteristics. Only values for features with $\kappa > 0$ are reported.

EC.10. ML versus Human

(a) Kidney

(b) Heart

(c) Liver

Figure EC.2 Relative frequency of reported drivers of 30-day readmission risk perception based on the survey responses. Acronyms are defined at the notes of Table EC.2.

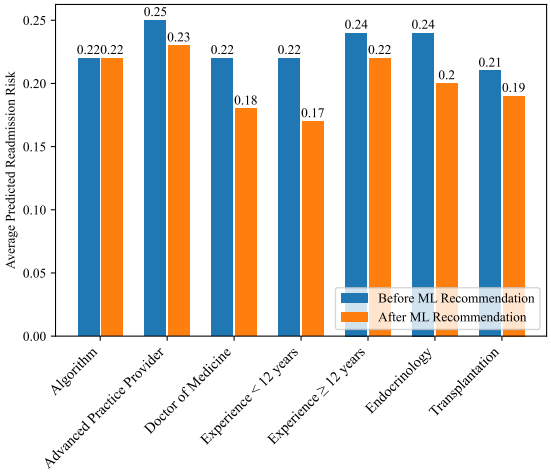
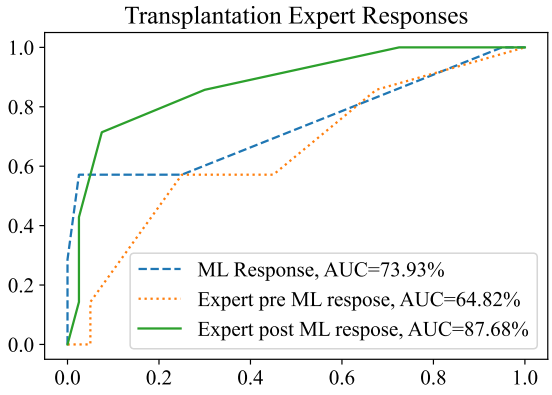


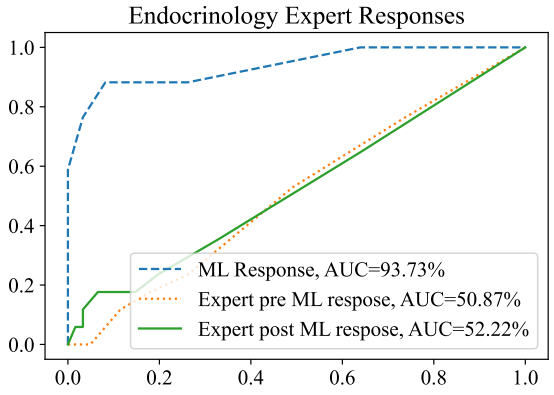
Figure EC.3 Human experts' risk perception as a function of provider heterogeneity.

Kidney	**Liver**	**Heart**
Creatinine at discharge ($p \leq 0.0001$)	Diabetes 2 history ($p \leq 0.001$)	HLA mismatch level ($p \leq 0.01$)
Recipient Age ($p \leq 0.0001$)	Average BG Value ($p \leq 0.001$)	Organ ($p \leq 0.01$)
Diabetes 2 history ($p \leq 0.0001$)	Insulin regimen ($p \leq 0.001$)	Average BG Value ($p \leq 0.01$)
Delayed graft function ($p \leq 0.0001$)	Donor BMI ($p \leq 0.01$)	Functional status at transplant ($p \leq 0.05$)
Insulin regimen ($p \leq 0.0001$)	Hyperglycemia ($p \leq 0.05$)	Creatinine at discharge ($p \leq 0.05$)
Average BG Value ($p \leq 0.001$)	Creatinine at discharge ($p \leq 0.05$)	Functional status at listing ($p \leq 0.05$)
Recipient BMI ($p \leq 0.001$)	Functional status ($p \leq 0.05$)	Presence of LVAD ($p > 0.05$)
Dialysis time ($p \leq 0.05$)	HbA1c at admission ($p > 0.05$)	Deceased donor ($p > 0.05$)
HbA1c at admission ($p \leq 0.05$)	Max BG Value ($p > 0.05$)	Ischemic Time (Hours) ($p > 0.05$)
Recipient race ($p > 0.05$)	Donor age ($p > 0.05$)	Diabetes 2 history ($p > 0.05$)

Table EC.8 T-test p-values for the ten most frequently selected risk factors for each organ that influenced clinical risk perception.

EC.11. Human or Algorithm?





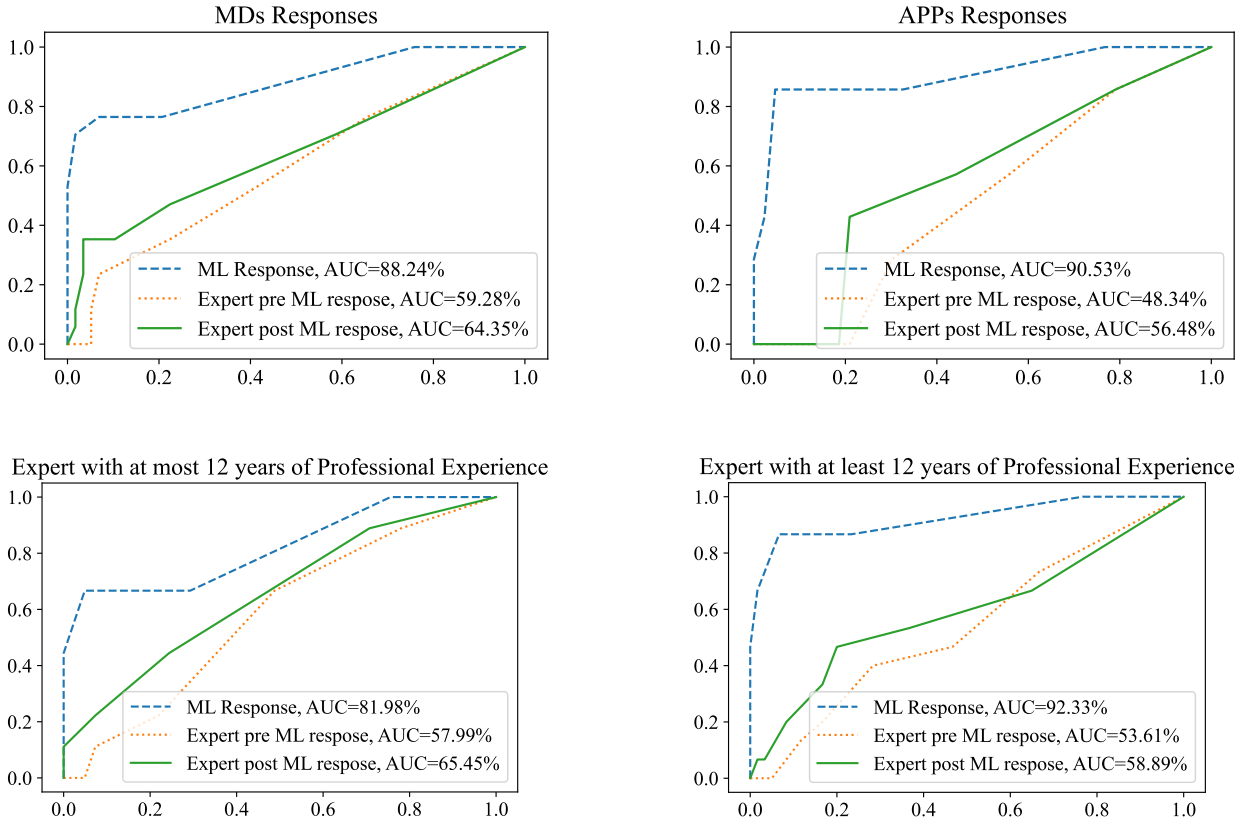


Figure EC.4 Survey Responses ROC Curves.

EC.12. Centaur Economic Value

Action Category	Advanced Practice Provider (%)	Endocrinology (%)	Doctor of Medicine (%)	Transplantation (%)
Nothing	30.00	38.46	46.67	42.55
Improve Glycemic Control	32.00	25.64	18.67	21.28
Schedule Early Follow-up	2.00	7.69	12.00	8.51
Treatment Education	30.00	20.51	8.00	10.64
Close Organ Monitoring	6.00	3.85	4.00	6.38
Ensure Caregiver Support	0.00	0.00	5.33	8.51
Extend Hospital Stay	0.00	3.85	5.33	2.13
Total Participants (N)	50	78	75	47

Table EC.9 Summary of survey responses to the question: What would you change in patient care during the index admission if the patient was at high risk upon discharge? Percentages are shown for each action category and total participant counts (N) for each provider/specialty.

EC.13. General Breakdown of ML versus Practitioner

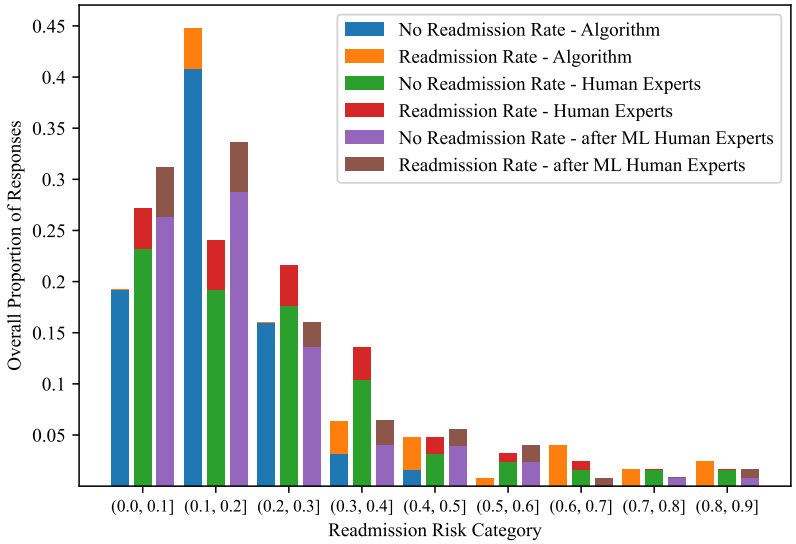


Figure EC.5 Illustration of the overall proportion of algorithm and physician responses in each risk category.