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

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

Karen Song
Harvard College, Harvard University, Cambridge, MA 02138, USA, karensong@college.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

There is growing evidence that machine learning (ML) algorithms can be used to develop accurate clinical risk scores for a wide range of medical conditions. However, the degree to which such algorithms can affect clinical decision-making is not well understood. Our work attempts to address this problem, investigating the effect of algorithmic predictions on human expert judgment. Leveraging a survey of medical providers and data from a leading U.S. hospital, we develop an ML algorithm and compare its performance with that of medical experts in the task of predicting 30-day readmissions after transplantation. We find that our algorithm is not only more accurate in predicting clinical risk but can also positively influence human judgment. However, its potential impact is mediated by the users' degree of algorithm aversion and trust. We show that, while our ML algorithm establishes non-linear associations between patient characteristics and the outcome of interest, human experts mostly attribute risk in a linear fashion. To capture potential synergies between human experts and the algorithm, we propose a human-algorithm "centaur" model and a framework to evaluate its performance in a real-world system. We show that our centaur model can outperform human experts and the best ML algorithm by systematically enhancing algorithmic performance with human-based intuition. Our results suggest that a centaur-enhanced risk assessment process that combines the power of human intuition with ML-based predictions could yield significant cost savings in practice by reducing readmissions through more accurate identifications of patients at risk of return at the time of discharge.*

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

---

* The study was approved by both Harvard's and Mayo Clinic's Institutional Review Boards. The authors gratefully acknowledge the help of three medical collaborators (Heidi Kosiorek, James Gilligan, and Janna Castro) who helped with data extraction and recruitment for the survey. The second author (Saghafian) acknowledges NSF Grant CMMI-1562645 "Data-Driven Management of Post-Transplant Medications" which partially enabled this work as well as funding and support from Harvard Data Science Initiative and Harvard Mossavar-Rahmani Center for Business and Government.

1. Introduction

Machine Learning (ML) is expected to transform the nature and delivery of healthcare (Rajkomar et al. 2019). Leveraging vast amounts of available data from various sources, 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 (Qayyum et al. 2020, Beam and Kohane 2018). Especially in the fields of radiology and computer vision, data-driven algorithms have been particularly successful (Rajpurkar et al. 2017). An increasing number of studies show that they can improve physician performance in complex tasks, such as tumor detection and diagnosis of cancer at an earlier stage compared to world-class specialists (Golden 2017, Yala et al. 2019). Yet, there is an "inconvenient truth" about ML in healthcare (Panch et al. 2019). Firstly, the vast majority of healthcare organizations, private or public, do not have the appropriate information system infrastructure to train and deploy ML models at the point of care (Sendak et al. 2019). Secondly, changing the interaction between physicians and patients with the introduction of data-driven models has encountered substantial challenges (Davenport and Kalakota 2019, Saghafian and Murphy 2021, Bertsimas and Orfanoudaki 2021).

"Algorithm aversion" lies at the center of some of these obstacles (Dietvorst et al. 2015). Formally defined, this term refers to the reluctance of human decision-makers to trust the recommendation of a data-driven algorithm, even if there is statistical evidence suggesting that it is more accurate than human experts (Jussupow et al. 2020). This phenomenon is even more prominent in the context of healthcare compared to other application areas due to the high stakes involved in medical decisions. As a result, the effectiveness of ML-based decision support tools in the clinical practice remains, to a large extent, a function of the behavioral characteristics of its users and their biases towards accepting algorithmic recommendations (Dai and Singh 2021).

Various studies have attributed the phenomenon of algorithm aversion in healthcare to the challenge of ML explainability (see, e.g., Babic et al. (2021)), which is often considered a major requirement for the successful implementation of ML models (Ahmad et al. 2018). Nevertheless, most well-established ML algorithms, including the celebrated neural networks, remain too complex to be directly understood by physicians (Tonekaboni et al. 2019). Even when an interpretable ML model is proposed, physicians may not always follow its recommendations, especially if it involves critical decisions that contradict their medical intuition. These observations raise the question of whether ML modelers need to directly incorporate the clinical thinking of human experts either a priori or a posteriori into the algorithm training and validation process. In particular, should clinical care move to a new model where a human-algorithm "centaur" improves the value delivered to patients? The idea of using this new model can be traced back to 2005, when a unique chess tournament, known as centaur chess, was introduced (see, e.g., (Goldstein et al. 2017), (Saghafian

2023b)). In it, humans and machines could collaborate together on the same team. In describing the results, the chess idol Garry Kasparov emphasized the following, which highlights the importance of utilizing the centaur model: "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). In this paper, we propose a human-algorithm centaur model of care delivery and investigate the implications of implementing it in a medical practice.

A different but related concept is known as human-in-the-loop, where researchers have proposed various ways to improve the performance of ML algorithms using human interactions (Xin et al. 2018). The centaur model differs from human-in-the-loop approaches mainly because centaurs are characterized by two important features: symbiotic learning and direct incorporation of human intuition (Saghafian 2023b). Mosqueira-Rey et al. (2022) identifies three broad types of learning that characterize the interactions between humans and ML algorithms: (a) active learning, in which the ML model maintains control of the process; (b) interactive ML, in which there is a tight interaction between humans and learning systems; and (c) machine teaching, where human domain experts guide the learning process. When ML is proven to be more accurate compared to human experts, it is unclear what type of interaction is likely to yield the highest benefit. This challenge becomes even more significant in the context of clinical care, where erroneous decisions can severely affect patients' lives.

Our work attempts to address these challenges by developing and implementing a centaur model. We investigate the performance of this centaur model in the context of solid-organ transplantations, focusing on preventing early hospital readmissions. To this end, we collaborate with physician experts at our partner hospital, the Mayo Clinic, and obtain a detailed clinical data set with information about more than 1,537 transplantations (see Table 1 for a data summary). We use this data set first to train and validate a ML algorithm capable of predicting the 30-day readmission risk post-transplantation. In parallel, we also design an interactive survey platform and utilize it to obtain physician experts' risk estimations based on the same observations that are seen by the ML algorithm. Using these data sources, we develop a human-algorithm centaur model and address the following research questions:

1. (Human or Algorithm): *Are humans or algorithms more accurate in predicting the risk of 30-day readmissions?*

2. (Algorithm Aversion): *Can ML algorithms influence human experts' risk estimations in the presence of algorithm aversion?*

3. (Reasoning and Risk Perception): *Do human experts and ML algorithms take into account the same clinical features in their risk estimations? Also, do human experts overestimate or underestimate the risk compared to ML algorithms?*

4. (Algorithm or Centaur): *Can a human-algorithm centaur model outperform the best ML algorithm, even when human experts are outperformed by such ML algorithms?*

5. (Centaur's Economic Value): *What are the economic gains stemming from implementing a centaur-enhanced risk assessment process in clinical practices?*

We contribute to both the management science and medical literature by providing answers to these research questions. Our contributions are six-fold:

- We develop and validate, to the best of our knowledge, the first successful ML algorithm for predicting 30-day readmission after transplantation of any of the major solid organs (kidney, liver, and heart). Our algorithm achieves an average out-of-sample Area Under the Receiver Operator Curve (AUC) of 84.0%.
- We demonstrate, using our survey platform, that a diverse group of human experts achieves significantly lower AUC (55.03%) compared to our ML algorithm (ceteris paribus, i.e., when provided with exactly the same information as the algorithm).
- Our study confirms the hypothesis that different clinical features drive physician decisions compared to the ML algorithm. For example, medical experts mainly focus on the history of diabetes and average Blood Glucose (BG) measurements, while our ML model primarily uses measures of BG variability. Our analysis also reveals that human experts overall overestimate the underlying risk compared to the ML algorithm. Moreover, patient and provider heterogeneity significantly affect the human experts' risk perception vis-a-vis the algorithm. Furthermore, we find that risk attribution by human experts can be explained to a very high degree by a linear regression model. This confirms past claims in the medical literature that humans tend to apply "linear" mental models of risk estimation. In contrast, our ML algorithm is highly non-linear.
- Our results show that, even though the practitioner's perception of risk is improved when informed about the ML recommendation, it remains weaker compared to the independent ML predictions. We demonstrate that this is because of algorithm aversion by showing that human experts put a low weight on the advice they receive from the ML algorithm.
- We show that when human intuition is systematically incorporated in the ML algorithm, the performance of the ML algorithm improves. Specifically, the AUC of the ML model improves by 2.42% when it is fed with insights from human experts. Thus, we demonstrate a human-algorithm centaur model, that is built by learning human intuition and feeding it into the ML algorithm, outperforms both the algorithm and the human experts, even though the human experts' performance is lower on average than the pure ML algorithm. This is because the centaur model takes advantage of the complementary effect of human intuition and the ML algorithm.

• We propose a generalizable framework for practitioners to train and evaluate the expected performance of centaur-enhanced human judgment processes in their practice, combining retrospective data from electronic health records and survey responses. In our specific medical application, we find that using the centaur-enhanced risk assessment process could significantly impact operational costs of clinical practices by better identifying patients at high risk of readmission and administrating more suitable interventions to prevent costly patient returns.

The remainder of the paper is organized as follows. Section 2 provides a summary of the literature relevant to our research questions. Section 3 introduces our best-performing ML model and presents the medical insights we derive from it. In Section 4, we describe the experimental study setting and present the design of our survey of medical experts. In Section 5, we compare the accuracy of our ML algorithm and human experts in predicting the risk of 30-day readmission. Section 6 focuses on the intuition behind the human expert responses and the reasoning behind their risk estimations. In Section 7, we introduce a generalizable human-algorithm centaur framework that allows us to augment the ML algorithm by directly incorporating the expert clinicians' intuition and measure its downstream performance when deployed as a decision support tool for human experts. In Section 8, we perform a simulation-based analysis guided by the survey responses to measure the expected economic value of implementing the centaur-enhanced risk assessment process in clinical practices. Finally, we conclude in Section 9 with an overview of the key findings.

2. Literature Review

Three main streams of literature are particularly relevant to our study: (1) empirical and theoretical studies that compare the performance of algorithms and humans as well as their perceptions of risk; (2) human-in-the-loop approaches that aim to augment algorithm recommendations with human guidance; (3) medical studies on 30-day readmission after solid-organ transplantation. In what follows, we briefly review each of these three streams.

An increasing number of studies suggests that supervised learning algorithms can lead to better estimations than humans across a wide variety of domains (He et al. 2015, Liu et al. 2018). Martin et al. (2004) showed that decision trees could outperform some of the most well-established legal experts in the country in predicting the outcomes of cases sent to the United States Supreme Court. In the context of healthcare, a recent review article identified nine studies from the medical field where the performance of a ML system was either on par with that of highly experienced clinicians or exceeded that of clinicians with less experience, focusing mostly on the areas of image recognition and deep learning (Shen et al. 2019). In the field of reinforcement learning, several algorithms have achieved superior performance compared to human experts, defeating the world's best players in cerebral games (see, e.g., Silver et al. 2018).

Our work directly complements the literature aiming to understand the process of human judgment and risk assessment (Skjong and Wentworth 2001). Hammond (1955) proposed the use of Brunswik (1952) lens model to capture with a linear function how individuals turn input from the task environment into clinical judgments. This work served as the foundation of "judgment analysis," validating across a multitude of settings that human experts tend to codify information in linear models when assigned with an estimation task (Brehmer and Joyce 1988, Karelaia and Hogarth 2008). Our study provides further evidence to this end, directly comparing human and algorithmic judgment in the context of healthcare and the effect that the latter could have on expert evaluations.

Yet, the degree and the factors that determine the impact of algorithmic recommendations on human decisions are still not very well understood. Psychology researchers have proposed various metrics to quantify the weight of advice related to human judgment in the context of algorithms (Harvey and Fischer 1997, Bailey et al. 2022, See et al. 2011). Applying such methodologies, Logg et al. (2019) provided evidence that people prefer algorithmic to human judgment. On the other hand, Yin et al. (2019) found that this finding is not universal. Their study claimed that people's trust in a ML algorithm depends on both the stated accuracy and its observed accuracy. Rudin and Ustun (2018) suggested that trust in ML systems in healthcare and criminal justice domains can only be gained through interpretable models, posing that the connection between humans and algorithms depends on the transparency of the latter. Wang et al. (2022) provided evidence that there might be conditions in which algorithmic transparency can be detrimental to strategic users, even if it is beneficial for the firm which deploys ML model. To provide further insights, Imai et al. (2020) developed a general-purpose statistical methodology that can experimentally evaluate the causal impact of algorithmic suggestions on human decisions. Kawaguchi (2021) found that humans are more likely to follow algorithmic recommendations when their forecasts are integrated into the algorithm. Finally, Saghafian (2023a) promoted a "two-way personalization" model, whereby incorporating preferences of physicians into a causal inference algorithm, treatment plans recommended by the algorithm are personalized to each patient and each physician.

Human-in-the-loop methodologies attempt to incorporate human feedback into the ML model deployment process (Wu et al. 2022). This field has predominantly focused on reinforcement learning settings, where expert guidance is particularly valuable at the initial stages of training (Amershi et al. 2014). In the context of healthcare, interactive and active ML may be particularly valuable in the presence of small data sets and high-risk decisions (Holzinger 2016). However, such approaches suggest a dynamic learning process between the human decision-maker and the algorithm. In the clinical practice, the current information system infrastructure often prohibits the baseline integration of the ML model into the clinical workflow (Panch et al. 2019). Thus, the expectation that

physicians will teach ML models over time may seem impractical. Artificial Intelligence systems may even affect the interactions between physicians. In a non-health context, Miklós-Thal and Tucker (2019) found that ML algorithms can impact the degree to which firms collude with each other in their pricing strategy. Ibrahim et al. (2021) introduced a system to elicit human judgment for prediction algorithms, assuming that experts have at their disposal subject information that is not available in the model input.

Arnold et al. (2019) conducted a prospective observational study comparing the predictive ability of an ML-based early warning system with physician judgment for clinical deterioration in hospitalized general internal medicine patients. Their analysis did not find significant differences between human experts and the AI system, but provided evidence that a ML model combining insights from both could lead to higher discrimination performance. Our work is related to this stream of literature, as it proposes a new framework to integrate expert advice into ML systems in the form of an exogenous predictive model that is trained on historical data of human judgment. Contrary to studies, such as Arnold et al. (2019), our method does not require an expensive prospective study since it only requires a limited number of expert responses in a survey that is based on retrospective data.

Our work complements the medical literature focusing on early readmissions (defined as occurring within 30 days after discharge) after solid organ transplantation (Li et al. 2016). Such readmissions constitute a costly and dangerous incident for both transplantation patients and hospitals that is often attributed to factors related to the index admission (Patel et al. 2016). Consequently, the reduction of these adverse events has become a key priority and an important quality measure for many hospitals and national health systems (Jencks et al. 2009). Improving quality measures, such as early readmissions, has become even more important for many hospitals in recent years, partially because public reporting of medical outcomes is being widely adopted by policymakers in an effort to increase quality transparency and improve the alignment between patients and provider capabilities (Saghafian and Hopp 2020).

Several studies have identified risk factors for either multiple or single readmissions using retrospective data and traditional statistical approaches, such as logistic regression (Schucht et al. 2020, Leal et al. 2017, Dols et al. 2018, Tavares et al. 2019). Haugen et al. (2018) differentiate their analysis for older and younger organ recipients while King et al. (2017) study adverse events like mortality and graft loss attributable to readmission after the transplant. The study of Covert et al. (2016) emphasizes the importance of patient understanding and adherence to medications as well as comorbidities, such as history of diabetes. Lubetzky et al. (2016) find that more than a quarter of early readmissions related to kidney transplanted patients could have been avoided with the use of continued outpatient management. Similar findings have been highlighted in the liver

and heart transplantation literature related to early readmission. However, the impact of donor characteristics seems to be more prominent for these organs compared to kidney (Chen et al. 2015, Yataco et al. 2016, Bachmann et al. 2018). In addition, Oh et al. (2018) stressed the importance of the length of stay during the index admission as well as the duration of warm ischemic time for liver transplantation patients. Zeidan et al. (2018) provided evidence that readmission rates can be reduced by improving access to outpatient services and hospital-local lodging for liver transplants in accordance with the findings of Lubetzky et al. (2016) for kidney transplanted patients.

Our work proposes, for the first time, a multi-organ early readmissions risk prediletion method after a transplantation, introducing one coherent model for kidney, liver, and heart transplant patients. We hypothesized that there are common patient factors across the three organs that drive the risk of early readmission. Specifically, we focused on the role of metabolic factors and the impact of BG management. Several studies have highlighted the importance of these variables during the immediate period after a transplant, uncovering commonalities between kidney and liver patients (Boloori et al. 2015, Chakkera et al. 2009, Munshi et al. 2020b, Werner et al. 2016). There is significant evidence that inpatient hyperglycemia can lead to future onset of diabetes mellitus (Chakkera et al. 2010, Munshi et al. 2021, 2020a) and targeted medication strategies are needed to avert potential adverse events for patients (Boloori et al. 2020, Saghafian 2023a). Orfanoudaki et al. (2023) provided evidence that early hospital readmission after a kidney transplantation is associated with glucometrics during the index admission. However, to the best of our knowledge, there has not been any study that directly explores the role of these factors across other types of solid organs and whether there are any commonalities between them in the context of early hospital readmissions.

3. Predicting Early Readmission after a Solid Organ Transplantation

Our analysis leverages retrospective clinical data obtained from electronic health records of the endocrinology and transplantation departments of the Mayo Clinic Arizona. Our data set comprises 1,537 de-identified cases of patients who received solid organ transplantation between September 25, 2015 and December 25, 2018. Only patients undergoing first-time solitary transplants were included in the study. Individuals who required readmission within the first 30 days following the index admission were identified using the hospital's operational records. We focused on the outcome of early readmission since it is a key measure indicating quality in transplantation care as well as a primary metric used by organizations such as the Centers for Medicare and Medicaid Services (CMS) for performance evaluation and reimbursement purposes. We supplemented this data with donor and organ-specific information from the United Network for Organ Sharing (UNOS) registry. The compiled patient vector summarizes all the information available in the hospital's electronic

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 1 Summary statistics of all clinical features for the patient population.

health records at the time of discharge. It includes both time-varying information (e.g., blood glucose values) as well as static variables from the hospital admission. Finally, as we will discuss in Section 4, we enhanced these data by running an independent survey of physician experts.

In what follows, we describe our patient population and the proposed ML model. Section 3.1 describes the clinical characteristics and the risk factors considered for our patient population. Section 3.2 outlines the training and validation process for our ML algorithm. In Section 3.3, we summarize the clinical insights that we gain from the ML model.

3.1. Patient Population

About 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. Table 1 summarizes the independent and dependent variables in the data. For numerical features, we report the mean value and the 95% confidence intervals. In the case of binary variables, we present the count and percentage of cases where the feature is prevalent. Our sample includes demographic information regarding both the donor and the recipient of the organ. To account for differences in the complexity of care at the hospital, Medicare Severity Diagnosis Related Group (MS-DRG) values were retrieved. We made use of the International Classification of Diseases, Tenth Revision (ICD-10) codes to determine which cases had a diagnosis of diabetes mellitus. To test whether metabolic factors affect the risk of early readmission, we incorporated multiple features, including average, minimum, and maximum values of the BG measurements (both hemoglobin A1c (HbA1c) and fasting plasma glucose levels) as well as the type of insulin regimen (basal, bolus, and combination) administered throughout the hospital stay. We report these metrics for the first, middle, and last 24 hours of hospitalization. Of note, 65.5% (16.9%) of the patients experienced hyperglycemia (hypoglycemia) during the index admission while 38.7% had history of diabetes. We incorporated organ-specific risk factors that we obtained from UNOS, although these variables contain information only applicable to a subset of organs or specific types of patients. Missing information was imputed using the MedImpute algorithm to account for temporal data associations (Bertsimas et al. 2021).

3.2. The Machine Learning Algorithm

To address the first research question we raised in Section 1, we begin our analysis by training multiple well-established ML algorithms to predict our outcome of interest (30-day readmission). Our goal is to derive one accurate and clinically relevant binary classification model that could assist physicians in assessing readmission risk and compare its performance with the existing process of human-based assessment. We compare the performance of logistic regression with regularization (to avoid overfitting), classification trees (CART), random forests, gradient boosted trees (XGBoost), support vector machines (SVM), and multi-layer perceptron (MLP) (Hastie et al. 2009, Breiman et al. 2017, Breiman 2001, Chen and Guestrin 2016, Cortes and Vapnik 1995, Rosenblatt 1958). 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 bootstrapped partitions of the data. We stratify the two sub-samples to ensure the same prevalence ratio of the outcome of interest. We conduct hyperparameter tuning using a bayesian optimization framework (Head et al. 2020) with the goal of maximizing the K-fold cross-validation AUC. We conduct the computational experiments in Python, leveraging the Scikit-learn library (Pedregosa et al. 2011).

Our results demonstrate that the XGBoost algorithm achieves superior performance compared to the other methods considered (see Table EC.1). Specifically, the XGBoost algorithm achieves a

mean 84.0% AUC on the testing set with 0.05% standard deviation. Our analysis suggests that AUC is higher (86.8%) for patients with a history of diabetes mellitus. The downstream performance of our ML models also significantly differs by 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 the lower AUC value for these patients. 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).

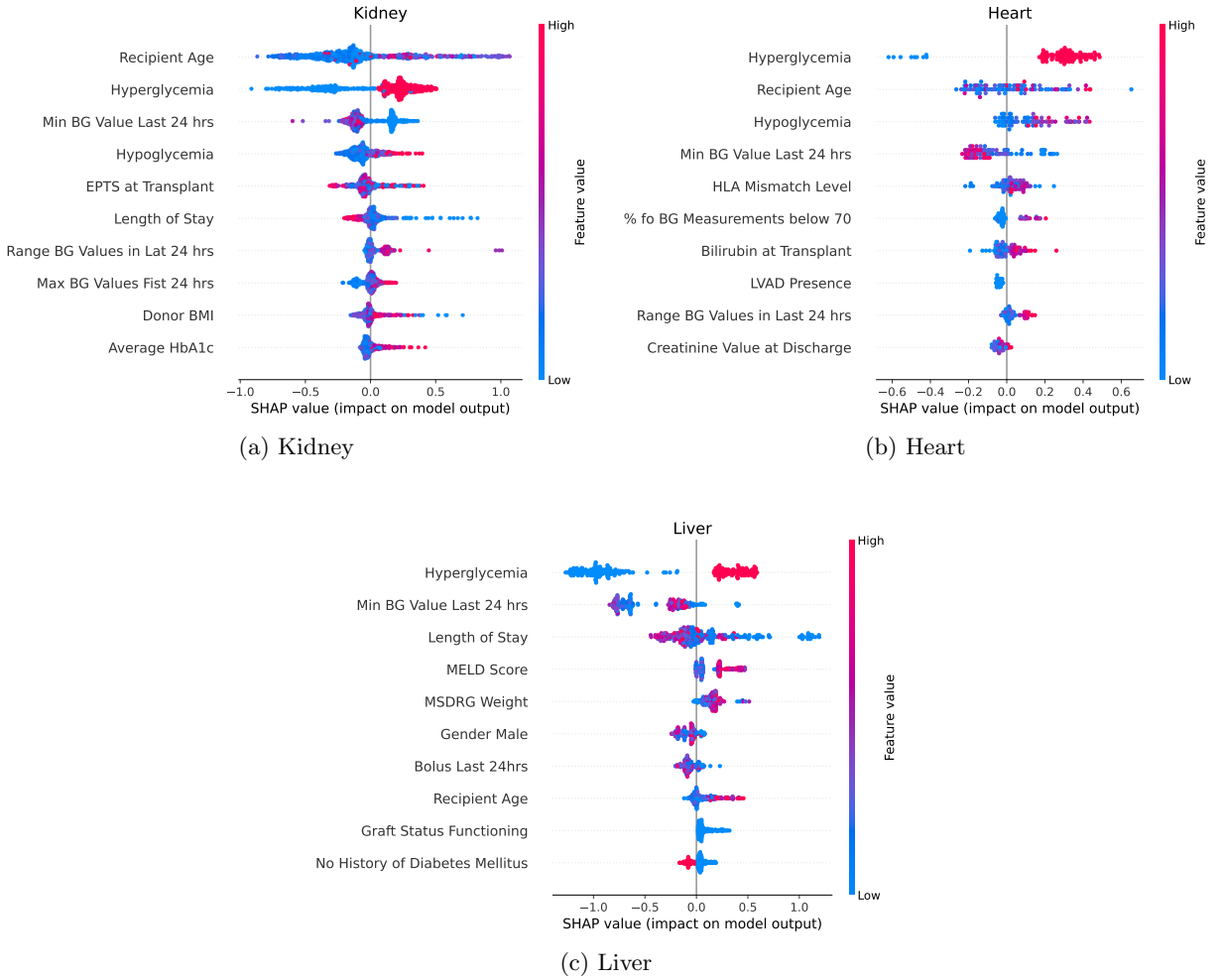
Finally, we observe that combining cases across all solid organs significantly improves the predictive accuracy for liver samples even though they form 23.7% of the overall data set. Excluding observations from any of the solid organs negatively affected the discrimination performance of the models. This finding provides evidence that there are common predictive pathways of risk that can explain the probability of readmission across the entire population of kidney, liver, and heart-transplanted patients.

3.3. Insights for Medical Practitioners

We use the SHapley Additive exPlanations (SHAP) framework (Lundberg and Lee 2017, Lundberg et al. 2020) to derive clinically relevant insights for medical practitioners from the best-performing ML algorithm (XGBoost). Our goal is to identify the main independent variables that can predict early hospital readmission per organ type and subsequently compare them with what human experts take into account when making the same predictions (see Section 6). Thus, we address the third research question we raised in Section 1. In addition, we use this tool to test our hypothesis of whether metabolic factors are associated with worse patient outcomes.

SHAP plots allow us to estimate the contribution of each variable to the predicted risk in the form of a normalized score between -1 and 1. The method leverages a game theoretic approach to approximate the XGBoost output with a linear model and estimate the average effect of each risk factor. Figure 1 highlights the 10 most important features of each type of organ. They are ordered by decreasing significance. Higher feature values are colored in red and lower feature values are in blue. Positive SHAP values are positively correlated with a higher chance of 30-day readmission and negative values indicate reductions in the risk of requiring additional hospitalization. We provide a comparison between the SHAP values and the coefficients and t-tests values of the regularized regression model in Table EC.2.

Our analysis validates our hypothesis, demonstrating that glucometrics are highly predictive of early hospital readmission after solid organ transplantation. Specifically, we find that the presence



(a) Kidney (b) Heart

(c) Liver

Figure 1 SHAP Plots for the proposed XGBoost model summarizing the risk contribution of the ten most important features per organ type. Acronyms are defined in the notes of Table 1.

of hyperglycemia is one of the two most important risk factors across all three organs. Our results emphatically highlight that not only high values but also abnormally low values of BG metrics can lead to a high risk of readmission (hypoglycemia and minimum BG value during the last 24 hours). In the case of kidney and heart patients, we identify that a higher range of BG values, defined as the difference between the maximum and minimum value, during the last 24 hours of the index admission is associated with a higher probability of re-hospitalization. While history of diabetes mellitus is widely regarded as one of the most significant risk factors for post-transplantation complications (Cook and Chakkera 2019), it is included in the ten most significant features only in the case of liver transplantation. Our analysis shows that BG control during the initial hospital admission is more predictive of the future patient trajectory than past history of diabetes. To the best of our knowledge, this is the first study that establishes the association between glucometrics throughout the index admission and early hospital readmission across all solid-organ types of transplantation. While Orfanoudaki et al. (2023) first highlighted the impact of suboptimal glucose

metrics during hospitalization on readmission in the case of kidney transplantation, our study generalizes these findings to other major organs (heart and liver).

Our ML algorithm also uncovers the role of other risk factors, validating the findings from past medical studies. We find that shorter length of stay and higher recipient age are associated with a higher risk of readmission, confirming previous evidence from the literature (Shankar et al. 2011, McAdams-Demarco et al. 2012). In the case of kidney transplants, our experiments indicate that the Estimated Post Transplant Survival (EPTS) score at transplant as well as donor Body Mass Index (BMI) are highly predictive of early hospital readmission (Schaenman et al. 2019, Dols et al. 2018). For liver patients, the ML model assigns high importance to the Model for End-Stage Liver Disease (MELD) score, Medicare Severity-Diagnosis Related Group (MSDRG) weight, recipient gender, and functioning status of the graft (Yataco et al. 2016). Finally, in the case of heart transplantations, the model identifies the Human Leukocyte Antigens (HLA) mismatch level, the values of bilirubin and creatinine, and the presence of Left Ventricular Assist Device (LVAD) as highly predictive features (Kim and Kim 2020). To conclude, we establish a relationship between blood glucose control and early readmission across liver, kidney, and heart transplantation, highlighting for the first time a general risk association between glucometrics and the outcome of interest. Our model constitutes the first early readmission risk score that combines heart, liver, and kidney organs into one model. In addition, it showcases commonalities in the risk profiles of patients who received different transplants. This is a new finding for the medical community, as we outline in the Literature Review that can lead to future follow-up research focusing on the impact of blood glucose management with a prescriptive rather than a predictive model. Finally, since (a) we are the first to combine multiple organs in a single model, (b) we use new variables that capture time-variant glucometrics, and (c) the literature lacks clear and comparable ML models with similar goals to ours, we do not believe direct comparisons other existing models of post-transplant early readmission risk would be logical.

4. Survey Design

To address our research questions, we designed an online survey platform and invited medical experts from the Mayo Clinic to respond to a series of questions for individual patient cases that had been previously evaluated by the ML algorithm. To provide fair comparisons between the ML model and the human experts, we made sure the ML model and the survey participants were given the same information.

4.1. Questions

Each participant was asked to review up to five patient cases. For each patient, the survey summarized in an interactive interface (see Figure 2) all the relevant case information that was available

in the data set. The task for each participant was to review the patient data and submit an answer to the following questions: (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? For this last question, the ML model's exact prediction (risk of 30-day readmission for the specific patient under review) was provided to the participants, and their response was collected to investigate whether and how much they updated their risk perception. Once the participant submitted their response to all the questions for one patient, the survey would prompt the user to the next case.

4.2. 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.

4.3. 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.

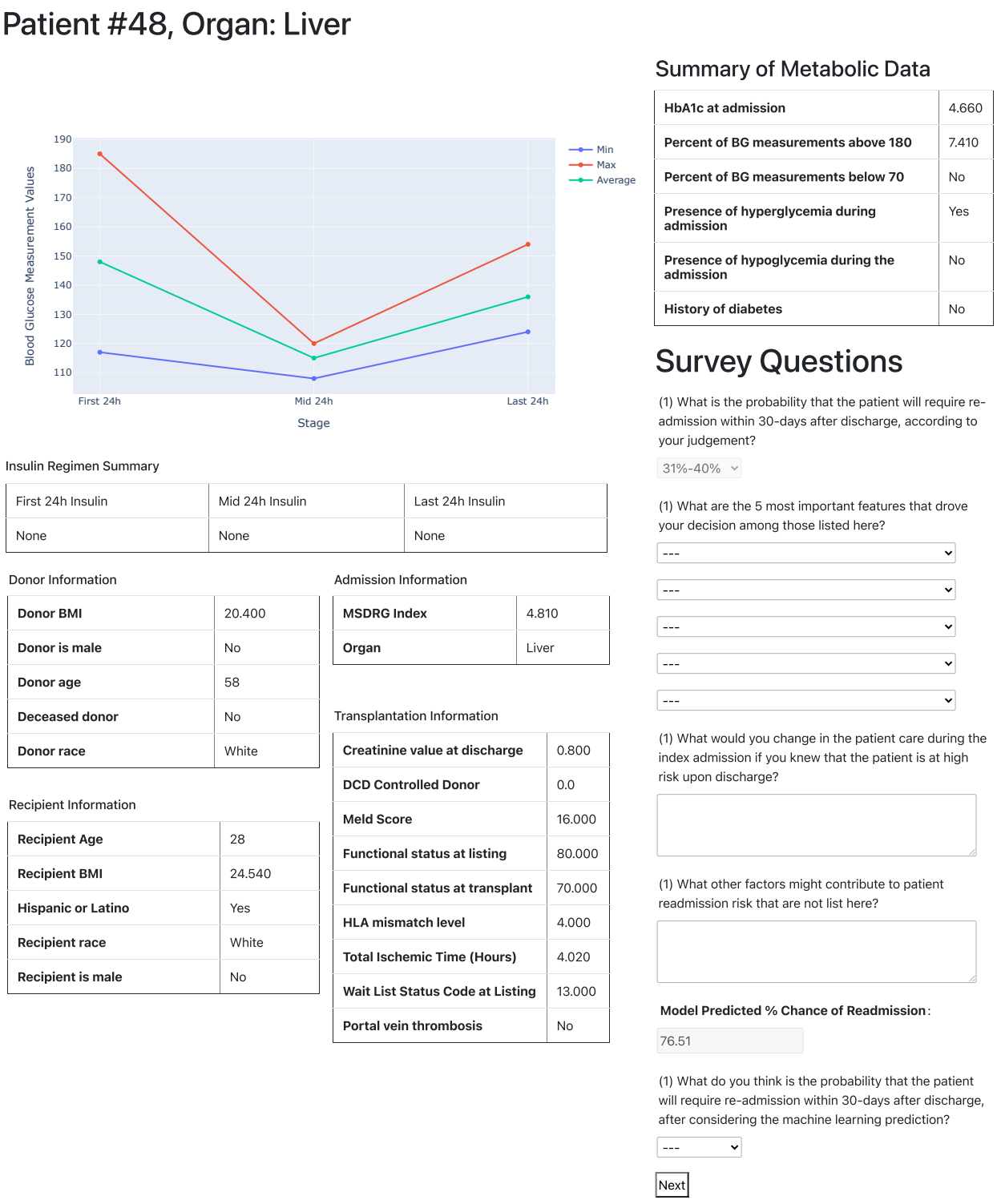


Figure 2 Illustration of the survey tool interface for an example liver patient.

4.4. 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 2. 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. 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), to ensure that their initial risk assessments and judgment are not affected. We programmed the user interface using the Django library in Python (Forcier et al. 2008).

5. Human or Algorithm? The Impact of Algorithm Aversion

In this section, we address the first two research questions we raised in Section 1. First, we show that a data-driven algorithm is more accurate compared to human experts in the task of predicting 30-day readmission of solid-organ transplant patients. Subsequently, our analysis reveals that ML predictions can positively influence human estimations, improving the experts' risk estimation. Table 2 summarizes our findings.

In total, 83 unique patients were reviewed and 125 distinct evaluations were recorded. We validated that we had collected the minimum required number of responses to secure, at most, a 8.0% sampling error, using the probability sampling method proposed by Dillman (2011) (see Table 5.1 in Dillman 2011). 47.0% of the cases were reviewed by one practitioner and 53.0% by two distinct experts. We used the Cohen's Kappa statistic to measure the inter-rater agreement between participants (Cohen 1960). We found that the agreement rate between human experts increased from $\kappa = 0.126$ ($p < 0.01$) to $\kappa = 0.348$ ($p < 0.01$) between the first and the second time that they provided their risk estimation. This finding suggests that the prediction of the ML model led to a higher consensus among human experts regarding the patients' future trajectory.

Clinical Subgroup	Experts AUC without ML	Experts AUC with ML	Weight of Advice (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 2 Discrimination performance summary of clinical experts' evaluations on the task of 30-day readmission.

In answering the first and last survey questions (Q1 and Q5), participants were asked to provide their risk estimation using intervals with 10% increments (e.g., $[0\%, 10\%), [10\%, 20\%)$, etc.). Following evidence from past research, it is very challenging for human assessors to distinguish between continuous values of risk (Sawyer 1966). It is considered more effective to ask experts to assign a class of risk instead (Goldberg 1970). Thus, we refrained from asking the human experts to submit a continuous value and assigned to them discrete yet granular categories of risk. This decision reduced the mental load of the survey on the participants and the required time to submit a response, facilitating the collection of a larger number of responses. This choice also allowed us to measure the discrimination performance of the participants which was essential in order to address the first three research questions presented in Section 1. To estimate the resulting AUC of the responses, we considered for each category the midpoint of the interval as the point estimate of the participant. We grouped the ML predictions following the same process and similarly replaced the ML suggested value with the midpoint of the interval that it belonged to. A random sample of patients from the testing set was included in the study. The average AUC performance of the ML model on that population was 88.55%.

First, to answer our first research question, we report the average AUC of human experts prior to the introduction of the ML algorithm in the survey (Q1). Overall, we notice a striking difference between the AUC of the ML algorithm (88.55%) and the survey participants (55.03%). Table 2 stratifies these results by participant subgroup. We find that transplantation experts achieve significantly higher results (64.82%) compared to their peers in the endocrinology department (50.87%). Similar findings were highlighted in the survey as we contrasted the discrimination performance of MDs (59.28%) and APPs (48.34%) as well as experts with less than 12 years (57.99%) and at least 12 years of professional experience (53.61%). We tested the statistical significance of the differences in the AUC performance between the ML model and the experts with and without the proposed algorithm's recommendations. We found that all differences were statistically significant with p-values< 0.001.

To address our second research question, once the ML model's evaluation was introduced in the survey, practitioners were asked to reconsider their risk estimations and submit a new answer (Q5). The updated responses were associated with an overall 11.28% relative improvement in AUC. We notice that the ML estimations positively influenced the survey participants across all clinical subgroups. We measure the degree of influence using the WoA metric (Harvey and Fischer 1997), which is measured as:

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

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 percentage of observations that met the latter exclusion criterion are included in parentheses in the fourth column of Table 2. The WoA metric reflects the degree to which clinicians weigh the algorithm's advice. Thus, it inversely relates to the extent of algorithm aversion and discounting (Yaniv 2004). If the final estimate is equal to the initial (ML) estimate, then WoA will be equal to 0 (1).

Using the WoA measure, we observe that transplantation experts and MDs are associated with the highest WoA (54.12% and 43.04% respectively). This is reflected in the relative AUC improvement of the former (35.26%) but less so in that of the latter (8.55%). In fact, APPs achieve double relative improvement (16.84%) compared to MDs with substantially lower WoA (26.36%). We do not identify significant WoA differences between medical practitioners with many or few years of professional experience. The detailed AUC curves for the ML model, as well as the survey participants both before and after the inclusion of the algorithm's estimation, are presented in Figure EC.1. When comparing individual performance with the WoA, we do not find a statistically significant correlation. This finding suggests the degree to which the ML suggestion impacts human experts is not associated with their baseline accuracy. This result can be explained by the fact that experts do not have access to information about the ground truth.

Put together, our results show that, when provided with exactly the same information, human experts are less accurate compared to the ML algorithm. However, 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. Specifically, for cases where $|\text{WoA}| \leq 1$, the AUC increases on average by 8.6% between the first and the second round of physician responses. However, when $|\text{WoA}| > 1$, the average absolute improvement in AUC is as high as 30%. Thus, our survey highlights participants with less algorithm aversion as a key barrier for some of the human experts in our study to achieve better performance.

6. Reasoning and Risk Perception

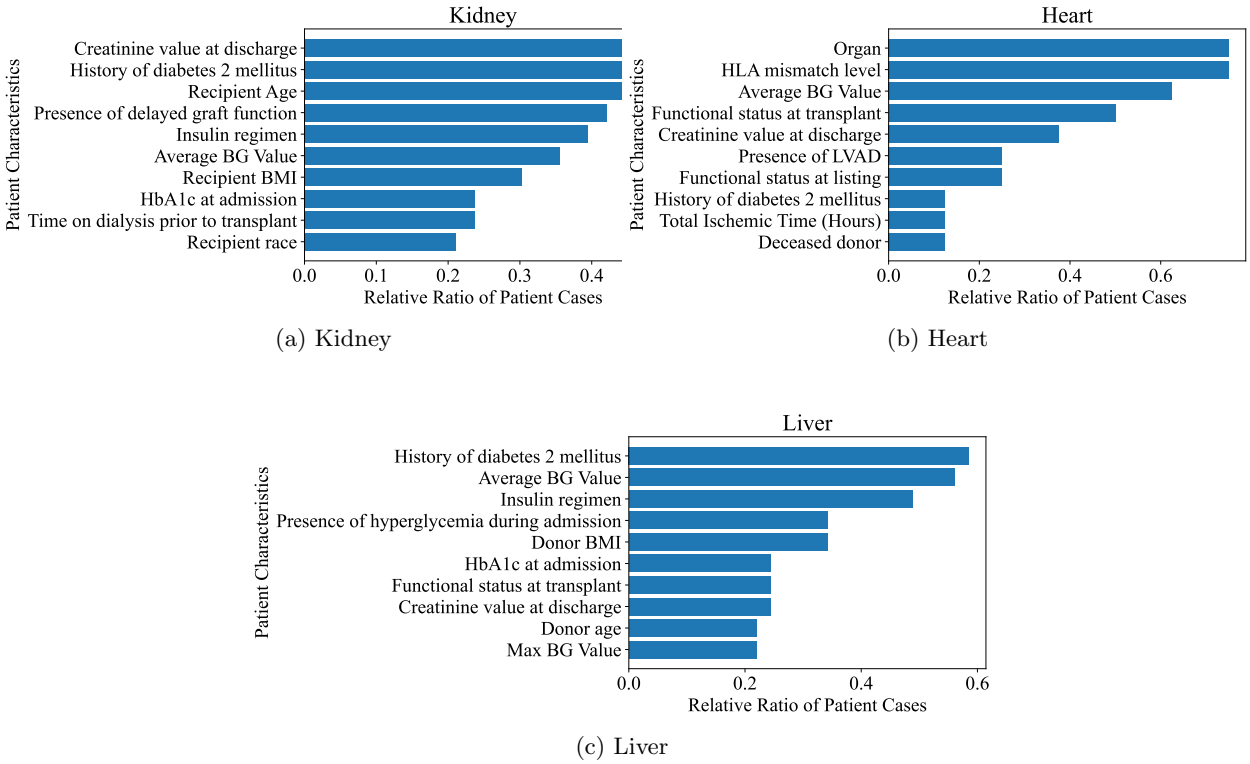
In this section, we aim to focus on the third research question we raised in Section 1: whether (a) human estimations are driven by the same clinical features as the ML model, and (b) human experts overestimate (or underestimate) readmission risk compared to the ML algorithm. We will investigate these in three ways. First, we report the clinical characteristics that survey participants identified as the primary drivers behind their risk estimations. Subsequently, we focus on the risk

perception of human experts and perform additional analyses to better understand the conditions under which they overestimate and underestimate the risk vis-a-vis the ML algorithm. Third, we employ a data-driven approach and develop linear regression models that estimate the survey responses directly from the patient characteristics.

6.1. Reported Clinical Drivers of Risk

The survey participants were asked to report five main patient characteristics that drove their risk estimation. To answer our third research question, we set the goal of uncovering the perceived drivers of readmission risk from medical practitioners and comparing them with those of the ML model. Figure 3 summarizes the survey's responses for each organ type. We report the respective p-values in Table EC.7. In more than 40% of kidney patients, experts identified the average creatinine value at discharge, history of diabetes, the recipient's age, and the presence of delayed graft function as one of the key factors determining their decision. Regarding metabolic information during the patient stay, practitioners distinguished the role of the average BG values (measured as fasting plasma glucose levels), the HbA1c (hemoglobin A1c) values at admission, and the type of insulin treatment. Moreover, the recipient's BMI and race as well as the time the patient spent on dialysis before the transplant were also included in 20% to 30% of the responses. In the case of heart transplantations, the type of organ and the HLA mismatch level were selected in more than 70% of patients. Similarly to kidney patients, the average BG value during the stay and creatinine value at discharge were reported in at least 40% of the survey submissions. The functional status at transplant (liver) and listing (heart), the total ischemic time (heart), and the presence of LVAD (heart) were organ-specific factors that significantly affected the experts' decisions. Finally, history of diabetes was selected substantially less often compared to kidney (15% of cases). The experts that looked into liver transplantations also highlighted history of diabetes and average BG values as the most important risk factors. They included insulin regimen as one of the primary drivers of their evaluation more often (close to 50% of the cases) in addition to donor BMI and age. Contrary to the other solid organs, the presence of hyperglycemia and maximum BG measurements were indicated as key factors that influenced their decision.

Comparing these findings with those in Section 3.3 reveals a stark difference between the key independent variables that influence the estimations of the ML algorithm compared to the human experts. The ML algorithm 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. On the other hand, medical experts identified across all organs the mean BG value



(a) Kidney (b) Heart

(c) Liver

Figure 3 Relative frequency of reported drivers of 30-day readmission risk perception based on the survey responses. Acronyms are defined at the notes of Table 1.

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).

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 in Appendix EC.3). For example, in more than 50% of the patient cases reviewed, the presence or absence of a history of diabetes was one of the five selected variables by the participants. Nevertheless, we observed limited agreement between providers that reviewed the same case ($k = 0.094$). These findings suggest that, while human experts often pick the same risk factors in the survey, they are not necessarily in a high degree of agreement when reviewing the same patient.

6.2. Overestimation versus Underestimation

Our findings in Section 3 are in line with long-standing literature in medical decision-making and hospital operations, which suggests that professionals should be given the flexibility to deviate from recommended protocols based on their perception of underlying risks. Recent studies show that such deviations can negatively affect outcomes such as the 30-day readmissions (see, e.g., (Atkinson and Saghafian 2022)), and hence, it is important to understand when professionals misperceive risks. In this section, we dig deeper and further investigate the misperception of risk among providers. Specifically, we fully address the answer to our third research question, we distinguish between two types of risk misperception: overestimation and underestimation. In particular, we focus on understanding whether and when human experts overestimate or underestimate risk as a function of true patient outcomes. This, in turn, enables us to observe whether there are specific situations in which human experts' intuition is more accurate than the ML algorithm.

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

In Table 3, we present the proportion of cases where human experts were over- and under-estimating the risk of readmission. In the first question (Q1), in which providers did not know the ML estimation, 4.17% (14.5%) of the participants agreed with the algorithm's recommendation for cases of readmission (no readmission). Overall, when human experts underestimated (overestimated) the risk, 66.67% (62.38%) of the cases were (not) associated with a hospital readmission. These findings are supported by the low out-of-sample AUC of the provider's estimations (see Section 5). Once survey respondents were prompted with the ML suggestion (Q5), their revised estimation improved. The agreement rate significantly increased to 22.77% (8.33%) for cases of no readmission (readmission). In addition, the relative frequency of underestimation and overestimation reduced as medical providers better calibrated their responses. Of note, in 33.33% of patients who required readmission, human experts had more realistic predictions regarding the patient trajectory and estimated higher risk than the algorithm. This shows that even though human experts

were worse at discriminating between the two outcomes, there are many individual patient cases where the human experts outperformed the algorithm. Thus, there could be potentially valuable human insights, and thus directly benefiting from them and feeding them to the ML algorithm might enhance its performance. To supplement this analysis, in Figure EC.2 we investigate the average predicted risk for different provider subgroups (see Section EC.3). Our results indicate two insights: (1) human experts, on average, overestimate the underlying risk (i.e., they are more conservative) than the ML algorithm, and (2) this overestimation behavior is consistent across different subgroups.

Put together, we find that although human experts tend to overestimate risks on average, there are situations in which their intuition allows them to outperform the ML algorithm. For this reason and to address our fourth research question, in the next sections, we first test if a data-driven model can be developed to learn the human experts' intuition. We then investigate whether creating a human-algorithm centaur by feeding the learned model to the ML algorithm can provide results superior to both the ML algorithm and the human experts.

6.3. Learning the Human Experts' Intuition in Risk Prediction

In this section, our goal is to develop a model that learns and accurately estimates the human experts' intuition in predicting the risk. To this end, for each patient $i \in [N]$ we consider a vector of features $\mathbf{x}_i$ as well as a binary outcome $y_i \in \{0,1\}$ indicating whether the patient was readmitted to the hospital within 30-days after discharge. We split the sample population in the training set T and the testing set E. We use well-established ML algorithms to derive a model $f(\mathbf{X})$ that estimates the binary outcome $\mathbf{y}$. The online survey comprises a subset of the testing set $S \subset E$. We let $R = E \setminus S$ be the set of patients that are part of the testing set but were not included in the survey. Let $\mathbf{z}_j$ be the vector of human expert characteristics that we record for each practitioner that participating in the survey $j \in [J]$. Each survey response $k \in [K]$ is associated with a patient vector $\mathbf{x}_k$, physician characteristics $\mathbf{z}_k$, a response $r_k \in [0,1]$ to the first question (Q1), and a response $w_k \in [0,1]$ to the last survey question Q5 (see Section 5). Our goal is to derive models that accurately characterize the experts' risk perception of hospital readmission, capturing the patient and human heterogeneity. Specifically, we train a model $g(\mathbf{X}, \mathbf{Z}) = \hat{\mathbf{r}}$ to estimate the response of the medical practitioners to the first survey question (denoted by $\mathbf{r}$) using all observations $[\mathbf{x}_k, \mathbf{z}_k] \in S$. In addition, we train a second model $h(\mathbf{X}, \mathbf{Z}, f(\mathbf{X})) = \hat{\mathbf{w}}$ to estimate the response of the medical practitioners to the last survey question (denoted by $\mathbf{w}$) leveraging all responses $[\mathbf{x}_k, \mathbf{z}_k, f(\mathbf{x}_k)] \in S$. Both models use as independent variables the patient information described in Section 3.1 and the expert's characteristics. The providers' responses to the first and fifth survey questions (Q1 and Q5 respectively) are used as the dependent variables. Given that some patients were reviewed by two

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

Table 4 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.

experts of different specialties, we developed a model that accounts for physician heterogeneity. As noted above, we consider each survey response as a distinct observation.

We assume both functions h and g are linear, and thus, we use ordinary least squares (OLS) regression to predict the continuous risk score provided by the survey participants (Hastie et al. 2009). We decided to focus on linear models as they were best able to capture the human experts' responses in the limited sample size of the survey. Non-linear models overfit the responses in the survey, obtaining low predictive power. In Section EC.4, we describe the comparative process that we followed to select the proposed linear models. Moreover, in Table EC.4 we summarize the predictive performance of the ML methods considered. In addition to the superior predictive performance, the OLS models allowed us to infer and characterize the primary independent variables that drive human risk perception.

We removed from the Q1 model all the independent variables with insignificant t-tests. For this reduced model, we examined the residuals for linearity, heteroscedasticity, auto-correlation, and outliers (Chatterjee and Hadi 2006). Due to the limited sample size, we trained the model on the entire population of survey responses (125 observations). The Mean Absolute Error (MAE) of the Q1 model was 0.1075 and the associated Brier score was 0.0204. The Bayesian Information Criterion (BIC) and Akaike Information Criterion (AIC) were -88.33 and -113.8, respectively (Akaike 1978, 1979). The linear regression coefficients, along with the resulting $p-$values of the $t-$tests, are summarized in Table 4.

The regression model confirms the human-reported drivers of risk presented in Figure 3. Specifically, it identifies that medical experts consider that higher recipient age at admission and BMI are associated with an increased probability of readmission. In addition, the data-driven approach

highlights the emphasis clinicians place on history of diabetes mellitus, HbA1c values at admission, and creatinine levels at discharge. We also uncover that while physicians report the average BG value as the primary driver of their risk evaluation (Figure 3), linear regression points to the average BG value in the last 24 hours rather than the entire stay. This perhaps highlights a potential bias that the BG dashboard (Figure 2) introduced to survey participants and potentially affected their perception of reported risk. The model also validates our hypothesis that the background of the medical providers may also affect the attribution of patient risk. While we did not find statistically significant differences between respondents from the endocrinology and transplantation departments, the regression model identified that APPs and experts with more professional experience are more conservative in their risk assessments.

The objective of the $h(\mathbf{X}, \mathbf{Z}, f(\mathbf{X})) = \hat{\mathbf{w}}$ model is to capture the expert's risk perception after being exposed to the ML recommendation. To be able to directly compare it with the $g(\mathbf{X}, \mathbf{Z}) = \hat{\mathbf{r}}$ model, we include the same set of independent variables. However, we include as an additional independent variable the ML recommendation that was shown to the practitioner. Thus, our model can capture the direct impact of the ML model on the human experts' risk perception, but also the difference in the relative importance of the remaining independent variables that describe the patient and practitioner characteristics. The MAE of the Q5 model was 0.1146 , the Brier score was 0.0252, the BIC and AIC were -74.18 and -102.5, respectively. The linear regression coefficients, along with the resulting $p-$values of the $t-$tests, are summarized in Table 4. The resulting model coefficient reveals the emphasis that the survey respondents placed on the ML recommendation. We observe that the ML suggestion becomes the most significant driver of human expert risk perception compared to both physician and patient characteristics. In addition, we find that the years of professional experience and history of diabetes 2 mellitus are not statistically significant in this model, while the importance of the practitioner's role increases. We do not observe other substantial changes to the remaining factors.

In addition to the OLS models, we replicate the same analysis using Bayesian learning (Gelman et al. 2013). Motivated by the literature suggesting that human perception might be better captured by Bayesian approaches (Knill and Pouget 2004), we trained three distinct models to learn systematically the human experts' response to the same questions (Q1 and Q5) (Salvatier et al. 2016). The resulting models did not show significant differences neither in the resulting expected value and standard deviation of the coefficients nor in the predictive performance for both downstream prediction tasks (see Table EC.5 in the Electronic Companion).

Put together, our results validate the hypothesis that the judgment of human experts can be captured by a linear model with very high precision validating the results of numerous studies from the psychology literature (Karelaia and Hogarth 2008, Brehmer and Joyce 1988, Cooksey

1996). Hammond (1955) first proposed the use of Brunswik (1952) model, posing that human judgement is modeled as a linear function of a set of cues that are shown to the decision maker. Since then, the psychology community has extensively studied the power of linear models as a way of capturing human ability in multiple-cue probability learning tasks or probabilistic concept identification tasks (Meehl 1954, Hammond et al. 1964, Tucker 1964, Castellan Jr 1973). Our results are aligned with these studies, indicating that the intuition of human experts in predicting risks can be sufficiently learned via a linear regression model. In the next section, we illustrate how feeding this learned regression model to the original ML algorithm can help us move towards a human-algorithm "centaur" model superior to both humans and the ML algorithm.

7. Algorithm or Centaur?

In this section, we address the fourth research question raised in Section 1. Specifically, leveraging the learned model of human intuition discussed in the previous section, we investigate whether we can improve the ML algorithm's performance by combining it with the power of human intuition, thereby creating a human-algorithm "centaur" model that can outperform both the human experts and the ML algorithm. We also move beyond our specific application and introduce a generalizable framework that enables the development and evaluation of centaur models that can be implemented in various practices.

7.1. The Centaur Model

In Section 6, we showed that a linear model could accurately and consistently capture the medical experts' intuition and risk perception. On average, the downstream AUC of medical providers on the outcome of interest is worse compared to the data-driven XGBoost model. However, our analysis identified cases in which humans more accurately estimated the future patient trajectory (see Table 3). Although these are isolated cases, we hypothesized that by integrating the human risk perception into the ML model, its performance could further improve. To investigate this, we updated the linear regression model proposed in Section 6.3 to remove the component of reviewer heterogeneity. That is, we train the model $l(\mathbf{X}) = \hat{\mathbf{r}}$ based on the survey responses $(\mathbf{X}_S, \mathbf{r}_S)$. The expected value of the predicted risk is $\sum_{i \in T} \frac{1}{|T|} l(\mathbf{x_i}) = 23.0\%$ and standard deviation of 18.0%. This model allows us to derive a function of human intuition that only depends on the patient characteristics (available in the training data), paving the path for a generalizable method of creating a centaur with a performance independent of the specific characteristics of survey respondents in our setting. Excluding physician heterogeneity from the model also allows us to directly apply this learner to all patient observations in our dataset (even those which were not part of the survey) and perform a fair comparison between the original ML model $f(\mathbf{X})$ and the centaur. We note, though, that removing the expert's characteristics from the model had a minor impact on the performance

of $l(\mathbf{X})$ compared to $g(\mathbf{X}, \mathbf{Z})$. The MAE increased to 0.1111 and the Brier score to 0.0219. The updated regression coefficients can be found in Table EC.6.

Subsequently, to derive the centaur model, we apply the resulting model $l(\mathbf{X})$ to the training set ($\mathbf{X}_T$) and use its output as an additional independent variable for the downstream ML algorithm prediction. By adding this value to the training data, we augment the feature space based on the hospital's health records with a composite variable that summarizes the experts' risk perception based on the same features. We then re-train the XGBoost algorithm using this enhanced training data and following the same process outlined in Section 3. Thus, we derive the centaur model $\hat{y}' = q(\mathbf{X}, l(\mathbf{X}))$, and use it to predict the true readmission status of patients $\mathbf{y}_T$. We split the data into the same partitions as before to derive a fair comparison of predictive performance. The average out-of-sample AUC of the centaur model across all random data partitions is 86.42% with a standard deviation of 0.16%. Thus, we observe that the performance improves by 2.42% compared to the original model, which did not include the experts' risk predictions, thus answering the fourth research question raised in the Introduction. This 2.42% is considerable, given that the original ML model already had high performance (best performance across a variety of ML models we had tested) and that increasing the AUC through any method is exponentially challenging once a high-performing AUC range is achieved.

Finally, it should be noted that this 2.42% improvement is an average value across three different solid organs (kidney, liver, and heart). However, by analyzing the performance across the three types of organs, we observe that the most significant benefit from the centaur model is manifested in the case of kidney and heart transplantations. For kidney transplantation, for example, we observe a 2.04% movement (from 77.0% to 79.04%). The AUC of the original XGBoost model for liver transplants was as high as 97.5%; thus, sustaining further improvement is a more challenging task than the other organs.

Put together, our findings provide empirical evidence that incorporating human experts' insights in the form of a systematic model can substantially improve algorithmic performance. These experiments show that a centaur, which incorporates human intuition into the algorithm, can improve the downstream performance of the algorithm even when the set of independent variables ($\mathbf{X}$) does not change. We hypothesize that the power of the centaur comes from the complementarity between human intuition and ML power. 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. 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 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.

7.2. A Generalizable Framework for Developing and Evaluating Centaurs

We now present a generalizable framework for developing centaur models that takes place during the derivation of the algorithm rather than the time of its deployment. Its prerequisite is an active study where human experts are required to provide their risk evaluation on the same task as the ML model, such as the one that we presented in the previous sections. This study can take place with only retrospective data, without significant changes and cost investments in IT infrastructure. Our framework leverages the output of the study to provide a scalable, inexpensive, and effective way to boost algorithmic performance without constant human supervision. Thus, practitioners could use our approach to evaluate the expected improvement that a centaur model could yield in their department in risk identification. In Section 8, we also show how to leverage the output of our framework to measure the economic value of the centaur. Recognizing the challenges and costs of ML integration in healthcare systems, this analysis could help decision makers better navigate the quality of care and financial trade-offs of a ML model implementation in clinical workflows. We present a summary of the proposed framework in Figure 4 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 (see Section 3.1).
2. **Split data into training and testing sets:** Following the standard procedure for ML training and validation, partition the data into a training set $(\mathbf{X}_T, \mathbf{y}_T)$ and testing set $(\mathbf{X}_E, \mathbf{y}_E)$ (see Section 3.2). The latter is further split into the subset of samples that will be used in the online survey $(\mathbf{X}_S, \mathbf{y}_S)$ and the subset of observations that will be part of the final model evaluation $(\mathbf{X}_R, \mathbf{y}_R)$.
3. **Train a baseline ML model:** Providing the observations $(\mathbf{X}_T, \mathbf{y}_T)$ as input, derive the best possible ML model using any vanilla supervised learning algorithm (see Section 3.2). The performance of the final model can be measured using the unseen samples of the entire testing set $(\mathbf{X}_E, \mathbf{y}_E)$.
4. **Design and conduct an online survey:** The next step is to design a survey to collect human expert responses similar to the one we used in our experiments (see Section 4). 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 $(\mathbf{Z}_S, \mathbf{w}_S, \mathbf{r}_S)$.

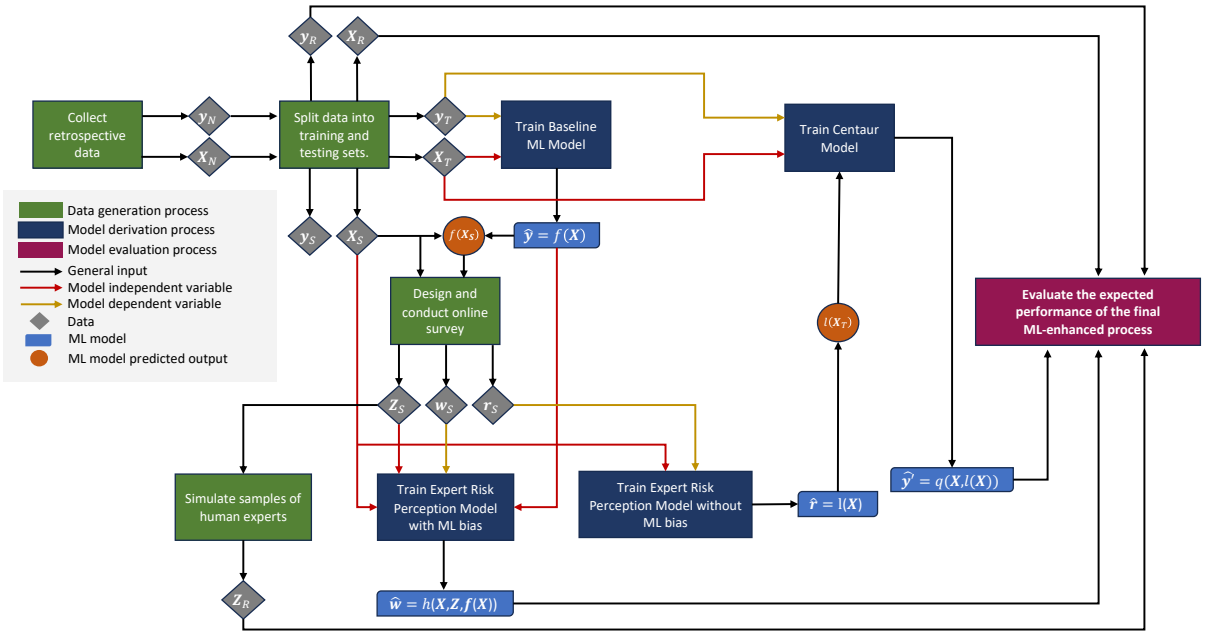


Figure 4 Summary illustration of the generalizable Centaur framework. We present in detail the flow of data, model training, validation, and evaluation that practitioners can follow to develop and deploy centaur models in practice.

5. **Train expert risk perception models without ML influence:** As presented in Section 7.1, 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.
6. **Train expert risk perception models with ML influence:** Similarly, following the model in Section 6.3, derive a second model $\hat{\mathbf{w}} = l(\mathbf{X}, \mathbf{Z}, f(\mathbf{X}))$, trained on the last series of survey responses $\mathbf{w}_S$) to codify the final risk perception of human decision makers with the use of the ML model while capturing the impact of human heterogeneity.
7. **Create a centaur model:** Using as input the predictions of the expert risk $l(\mathbf{X}_T)$ model with the original training set $(\mathbf{X}_T, \mathbf{y}_T)$, update the baseline ML model to derive a centaur model $q(\mathbf{X}, l(\mathbf{X}))$. The centaur model is able to combine human intuition with the algorithm insights into a single learner (see Section 7.1).
8. **Simulate samples of human experts:** To associate the observations of the testing set $[R]$ that were not part of the survey with human expert features, use the characteristics of the survey respondents to generate a representative sample of synthetic experts $\mathbf{Z}_R$. This simulated sample serves as a representative pool of the human decision-makers that work in the organization where the ML model will be deployed. Alternatively, if the survey collects responses for all patients in E, randomly partition the testing set in the subsets R and S to derive two independent population for the centaur and human experts model derivation and validation.

9. **Evaluate the expected performance of the model for implementation in practice:** As the last step, use the $\hat{\mathbf{w}} = h(\mathbf{X}, \mathbf{Z}, q(\mathbf{X}, l(\mathbf{X}))$ model, evaluated on the unseen portion of the testing set $(\mathbf{X}_R, \mathbf{y}_R)$ to measure the expected aggregate performance of the process where the human decision makers will receive advice from the centaur model $q(\mathbf{X}, l(\mathbf{X}))$.

Building the centaur framework discussed above is an inexpensive way for real-world organizations, not limited to healthcare, to incorporate the intuition of their own human experts into a baseline ML model, and estimate the value that this could bring to their organization. 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 in their system after deployment. This improvement can come with low cost for the organization as the only additional input required is the responses from a retrospective survey similar to the one conducted in our study. For more details about how we gauge the performance of our centaur model and apply steps 8 and 9 as presented above, we refer the reader to Section EC.7.

8. The Economic Value of the Centaur

We now answer the fifth research question we raised in Section 1: what is the economic value that our centaur approach could bring to the healthcare system? To answer this question, we asked survey respondents what actions they would pursue if they knew that a patient would require readmission. Table 5 summarizes the responses we received. We identified seven primary categories of action whose frequency varies depending on the role and specialty of the provider.

Our results demonstrate that in 40% of the cases, providers would not change anything in patient care. The rate is lower for APPs (30%) compared to MDs (46.67%). The most common response was no alteration in patient care. The second most popular course of action focuses on improving glycemic control. According to 24% of experts, effectively managing BG measurements could avert a potential re-hospitalization, especially for patients with a history of diabetes. We did not find significant differences between the endocrinology and transplantation teams in this regard, but we observed that MDs resort less often to that option compared to APPs (18.67% and 32.00% respectively). In addition to better control of metabolic factors, the endocrinology team and APPs placed a lot of emphasis on treatment education to ensure high adherence to post-transplantation and BG therapy. MDs favored close patient monitoring practices, such as extending the hospital stay, scheduling early and regular follow-ups, and continuous checks of the organ's health. Finally, transplantation physicians highlighted the importance of ensuring caregiver support at home. Especially in the context of elderly patients, the latter emphasized that early readmission could be avoided in the presence of high-quality home support after the surgery.

This analysis highlights the clinical and operational levers of action that transplantation centers could use to improve patient outcomes and reduce re-hospitalization rates. In addition, it illustrates

Action Category	**Responses %**	**Action Category**	**Responses %**
Advanced Practice Provider	$N = 50$	Endocrinology	$N = 78$
Nothing	30.00	Nothing	38.46
Improve glycemic control	32.00	Improve glycemic control	25.64
Schedule early follow up	2.00	Schedule early follow up	7.69
Treatment education	30.00	Treatment education	20.51
Close organ monitoring	6.00	Close organ monitoring	3.85
Ensure caregiver support at home	0.00	Ensure caregiver support at home	0.00
Extend hospital length of stay	0.00	Extend hospital length of stay	3.85
Doctor of Medicine	$N = 75$	Transplantation	$N = 47$
Nothing	46.67	Nothing	42.55
Improve glycemic control	18.67	Improve glycemic control	21.28
Schedule early follow up	12.00	Schedule early follow up	8.51
Treatment education	8.00	Treatment education	10.64
Close organ monitoring	4.00	Close organ monitoring	6.38
Ensure caregiver support at home	5.33	Ensure caregiver support at home	8.51
Extend hospital length of stay	5.33	Extend hospital length of stay	2.13

Notes. The providers' answers have been clustered into seven categories. The Table outlines the percentage of responses that belong to each category for the two types of providers and specialties considered.

Table 5 Summary of survey responses to the changes in care question: What would you change in the patient care during the index admission if you knew that the patient is at high risk upon discharge?

the degree to which clinical teams are willing to adapt their care practices and their differences based on the type of services they provide. As healthcare systems move towards value-based care, accurate risk scores for adverse event prediction will only be effective if they lead to changes in patient care for individuals at high risk. Thus, the integration of ML risk scores, such as the one that we present, should be accompanied by a mapping of options and associated operational processes that clinical teams could resort to at the time of discharge to avoid future adverse events. In our setting, our analysis highlighted five broader categories of action providers would be keen to follow: (1) review of BG treatment; (2) treatment education program; (3) early follow up with a transplantation or endocrinology expert; (4) extended stay at the hospital; (5) home care support.

In what follows, our goal is not to provide a causal effect estimation for the use of our centaur model (e.g., on the patient health outcomes), but rather conduct a rigorous analysis that will allow us to measure the economic value of the proposed ML-enhanced decision-making process. To this end, consider a risk threshold $\tau \in [0, 1]$ above which the medical providers at our partner hospital would use one of the above-mentioned interventions to avoid a potential hospital readmission. We assume that the expected cost of such interventions is WC on average and that the respective cost of readmission is RC. We acknowledge that these interventions do not have a deterministic effect on transplanted patients, and thus, only a p portion of those receiving an intervention will avoid

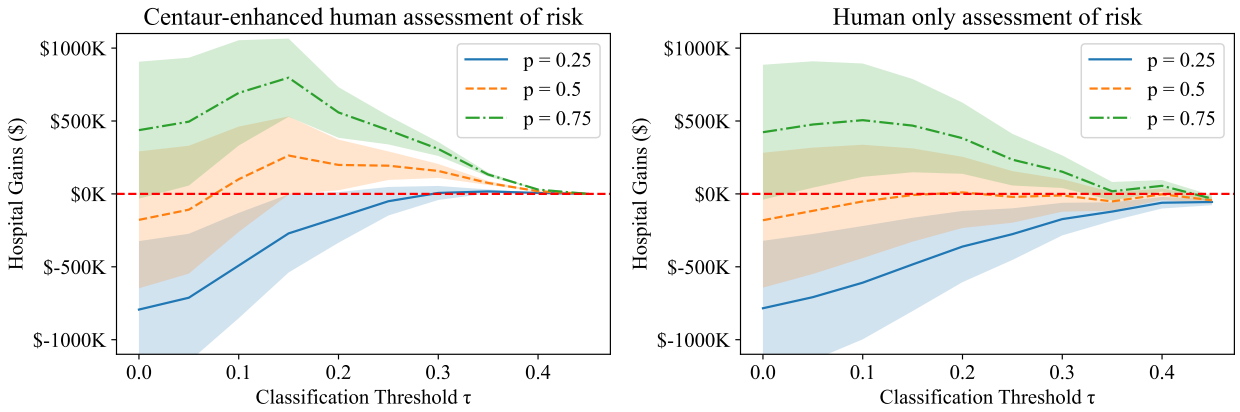


Figure 5 Comparison of estimated economic gains for the Mayo Clinic between the existing process of readmission risk evaluation that involves only human judgement and the proposed centaur-enhanced process where medical practitioners receive recommendations from the proposed centaur model.

readmission. Based on these assumptions, the annual economic value of any readmission prediction model m is:

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

where $RP_{\tau,m}$ is the expected annual number of patients with predicted risk higher than τ by model m who are re-admitted after receiving the intervention and $P_{\tau,m}$ is the expected annual number of patients with predicted risk higher than τ by model m.

We conduct a sensitivity analysis on the parameters p and WC, setting the value of $RC = \$27,000$ (Weiss and Jiang 2021). We compare the performance of two models (a human-only risk assessment model resembling the current practice and the proposed centaur-enhanced process) for different levels of the risk threshold τ, and characterize the potential cost savings of the centaur-enhanced risk assessment process. We let the intervention cost be $WC \in [\$2,000, \$4,000]$ following cost estimations from the literature (Mann et al. 2020, Serrano et al. 2019, Zimmerman et al. 2022). We assume that the average probability of success, p, lies in the set $\{0.25, 0.5, 0.75\}$ and that the classification threshold $\tau \in (0, 0.45]$. We summarize our findings in Figure 5.

Our analysis highlights the benefits that hospital administrators can achieve by implementing the centaur model. 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 approach 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, the hospital is unable to gain benefits with more conservative classification thresholds due to the high number of false positive cases.

Overall, 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. Since developing and testing the centaur model is fairly inexpensive, it can offer sufficient financial incentives to hospital administrators for implementation in their clinical practices.

9. Conclusions

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. However, by codifying human intuition into a predictive model, we propose a human-algorithm centaur model to bridge the gap between the two. We show that the centaur outperforms both the best pure ML predictions and those of human experts. This finding is partly driven by the fact that human expert intuition can complement the power of ML models. Thus, 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 this 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. To provide fair comparisons, we intentionally limited what was provided to human experts in our surveys to those available to the baseline ML algorithm. However, non-explicit knowledge may significantly affect clinical decisions during medical care at the hospital (Patel et al. 1999). For example, APPs and MDs often act upon visual signals or conversations during their contact with the patients that are not included in the electronic health records of the hospital and, thus, cannot be captured by the algorithm (Reinders 2010). 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 (CMS). Thus, our analysis did not consider longer-term horizons of potential readmission, such as a 60-day or 90-day window. Future studies could focus on extending our results to these post-transplant adverse events and other types of quality metrics, such as patient survival, length of stay, and graft survival. 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 and patient support at home, 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. There might be many reasons why

human-experts tend to over- or under-estimate risk compared to the ML model. Human perception of risk can be influenced by various factors, including over-confidence, fear of lawsuits, and specific shortcomings in medical training, among others. Studying what is driving the physicians experts' behavior is beyond our area of focus, and we leave it as a potential extension of this work to future studies. Finally, future research can also investigate adding more value to our proposed human-algorithm centaur framework by incorporating interpretable models as the baseline ML algorithm. Understanding whether there is some synergy between using interpretable models and our proposed method of directly capturing human intuition and feeding to a baseline ML algorithm can be invaluable.

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

Ahmad MA, Eckert C, Teredesai A (2018) Interpretable machine learning in healthcare. *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, 559–560.

Akaike H (1978) On the likelihood of a time series model. *Journal of the Royal Statistical Society: Series D (The Statistician)* 27(3-4):217–235.

Akaike H (1979) A bayesian extension of the minimum AIC procedure of autoregressive model fitting. *Biometrika* 66(2):237–242.

Amershi S, Cakmak M, Knox WB, Kulesza T (2014) Power to the people: The role of humans in interactive machine learning. *AI Magazine* 35(4):105–120.

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.

Atkinson M, Saghafian S (2022) Who should see the patient? On deviations from preferred patient-provider assignments in hospitals. *Health Care Management Science (forthcoming), available at SSRN* .

Babic B, Gerke S, Evgeniou T, Cohen IG (2021) Beware explanations from AI in health care. *Science* 373(6552):284–286.

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.

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

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.

Boloori A, Saghafian S, Chakkera HA, Cook CB (2020) Data-driven management of post-transplant medications: An ambiguous partially observable Markov decision process approach. *Manufacturing & Service Operations Management* 22(5):1066–1087.

Brehmer B, Joyce CRB (1988) *Human judgment: The SJT view* (Elsevier).

Breiman L (2001) Random forests. *Machine Learning* 45(1):5–32.

Breiman L, Friedman JH, Olshen RA, Stone CJ (2017) *Classification and Regression Trees* (Routledge).

Brunswik E (1952) The conceptual framework of psychology. *(No Title)* .

Castellan Jr NJ (1973) Comments on the "lens model" equation and the analysis of multiple-cue judgment tasks. *Psychometrika* 38(1):87–100.

Chakkera HA, Knowler WC, Devarapalli Y, Weil EJ, Heilman RL, Dueck A, Mulligan DC, Reddy KS, Moss AA, Mekeel KL, et al. (2010) Relationship between inpatient hyperglycemia and insulin treatment after kidney transplantation and future new onset diabetes mellitus. *Clinical Journal of the American Society of Nephrology* 5(9):1669–1675.

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.

Chatterjee S, Hadi AS (2006) *Regression Analysis by Example* (John Wiley & Sons).

Chen P, Wang W, Yan L, Yang J, Wen T, Li B, Zhao J, Xu M (2015) Risk factors for first-year hospital readmission after liver transplantation. *European Journal of Gastroenterology & Hepatology* 27(5):600–606.

Chen T, Guestrin C (2016) Xgboost: A scalable tree boosting system. *Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining* 785–794.

Cohen J (1960) A coefficient of agreement for nominal scales. *Educational and psychological measurement* 20(1):37–46.

Cook CB, Chakkera H (2019) Diabetes mellitus and renal transplantation. *Endocrine Disorders in Kidney Disease* 75–81.

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

Cortes C, Vapnik V (1995) Support-vector networks. *Machine Learning* 20(3):273–297.

Covert KL, Fleming JN, Staino C, Casale JP, Boyle KM, Pilch NA, Meadows HB, Mardis CR, McGillicuddy JW, Nadig S, et al. (2016) Predicting and preventing readmissions in kidney transplant recipients. *Clinical Transplantation* 30(7):779–786.

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`.

Davenport T, Kalakota R (2019) The potential for artificial intelligence in healthcare. *Future Healthcare Journal* 6(2):94.

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).

Dols JD, Chargualaf KA, Spence AI, Flagmeier M, Morrison ML, Timmons A (2018) Impact of population differences: Post-kidney transplant readmissions. *Nephrology Nursing Journal* 45(3):273–281.

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).

Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB (2013) *Bayesian data analysis* (CRC press).

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.

Golden JA (2017) Deep learning algorithms for detection of lymph node metastases from breast cancer: helping artificial intelligence be seen. *JAMA* 318(22):2184–2186.

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

Hammond KR (1955) Probabilistic functioning and the clinical method. *Psychological review* 62(4):255.

Hammond KR, Hursch CJ, Todd FJ (1964) Analyzing the components of clinical inference. *Psychological review* 71(6):438.

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

Hastie T, Tibshirani R, Friedman JH, Friedman JH (2009) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, volume 2 (Springer).

Haugen CE, King EA, Bae S, Bowring MG, Holscher CM, Garonzik-Wang J, McAdams-DeMarco M, Segev DL (2018) Early hospital readmission in older and younger kidney transplant recipients. *American Journal of Nephrology* 48(4):235–241.

He K, Zhang X, Ren S, Sun J (2015) Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *Proceedings of the IEEE International Conference on Computer Vision*, 1026–1034.

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* .

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.

Kim MJ, Kim K (2020) Unplanned readmission of patients with heart transplantation in 1 year: A retrospective study. *Journal of Advanced Nursing* 76(3):824–835.

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.

Knill DC, Pouget A (2004) The bayesian brain: the role of uncertainty in neural coding and computation. *TRENDS in Neurosciences* 27(12):712–719.

Leal R, Pinto H, Galvão A, Rodrigues L, Santos L, Romãozinho C, Macário F, Alves R, Campos M, Mota A, et al. (2017) Early rehospitalization post–kidney transplant due to infectious complications: Can we predict the patients at risk? *Transplantation Proceedings*, volume 49, 783–786 (Elsevier).

Li AHt, Lam NN, Naylor KL, Garg AX, Knoll GA, Kim SJ (2016) Early hospital readmissions after transplantation: burden, causes, and consequences. *Transplantation* 100(4):713–718.

Liu Y, Zhang H, Zeng L, Wu W, Zhang C (2018) Mlbench: benchmarking machine learning services against human experts. *Proceedings of the VLDB Endowment* 11(10):1220–1232.

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

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.

Lundberg S, Lee SI (2017) A Unified Approach to Interpreting Model Predictions. Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R, eds., *Advances in Neural Information Processing Systems 30*, 4765–4774 (Curran Associates, Inc.), URL `http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf`.

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.

Martin AD, Quinn KM, Ruger TW, Kim PT (2004) Competing approaches to predicting supreme court decision making. *Perspectives on Politics* 2(4):761–767.

McAdams-Demarco M, Grams M, Hall E, Coresh J, Segev D (2012) Early hospital readmission after kidney transplantation: patient and center-level associations. *American Journal of Transplantation* 12(12):3283–3288.

Meehl PE (1954) Clinical versus statistical prediction: A theoretical analysis and a review of the evidence. .

Miklós-Thal J, Tucker C (2019) Collusion by algorithm: Does better demand prediction facilitate coordination between sellers? *Management Science* 65(4):1552–1561.

Mosqueira-Rey E, Hernández-Pereira E, Alonso-Ríos D, Bobes-Bascarán J, Fernández-Leal Á (2022) Human-in-the-loop machine learning: A state of the art. *Artificial Intelligence Review* 1–50.

Munshi VN, Saghafian S, Cook CB, Aradhyula SV, Chakkera HA (2021) Use of imputation and decision modeling to improve diagnosis and management of patients at risk for new-onset diabetes after transplantation. *Annals of Transplantation* 26:e928624–1.

Munshi VN, Saghafian S, Cook CB, Steidley DE, Hardaway B, Chakkera HA (2020a) Incidence, risk factors, and trends for postheart transplantation diabetes mellitus. *The American Journal of Cardiology* 125(3):436–440.

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

Oh SY, Lee JM, Lee H, Jung CW, Yi NJ, Lee KW, Suh KS, Ryu HG (2018) Emergency department visits and unanticipated readmissions after liver transplantation: A retrospective observational study. *Scientific Reports* 8(1):1–9.

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.

Patel VL, Arocha JF, Kaufman DR (1999) Expertise and tacit knowledge in medicine. *Tacit Knowledge in Professional Practice*, 89–114 (Psychology Press).

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).

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, et al. (2011) Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12:2825–2830.

Qayyum A, Qadir J, Bilal M, Al-Fuqaha A (2020) Secure and robust machine learning for healthcare: A survey. *IEEE Reviews in Biomedical Engineering* 14:156–180.

Rajkomar A, Dean J, Kohane I (2019) Machine learning in medicine. *New England Journal of Medicine* 380(14):1347–1358.

Rajpurkar P, Irvin J, Zhu K, Yang B, Mehta H, Duan T, Ding D, Bagul A, Langlotz C, Shpanskaya K, et al. (2017) Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225* .

Reinders H (2010) The importance of tacit knowledge in practices of care. *Journal of Intellectual Disability Research* 54:28–37.

Rosenblatt F (1958) The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review* 65(6):386.

Rudin C, Ustun B (2018) Optimized scoring systems: Toward trust in machine learning for healthcare and criminal justice. *Interfaces* 48(5):449–466.

Saghafian S (2023a) Ambiguous dynamic treatment regimes: A reinforcement learning approach. *Management Science* .

Saghafian S (2023b) Effective Generative AI: The Human-Algorithm Centaur. *Available at SSRN 4587250* .

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, Murphy SA (2021) Innovative health care delivery: The scientific and regulatory challenges in designing mhealth interventions. *National Academy of Medicine Perspectives* 2021.

Salvatier J, Wiecki TV, Fonnesbeck C (2016) Probabilistic programming in python using pymc3. *PeerJ Computer Science* 2:e55.

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

Schaenman J, Castellon L, Liang EC, Nanayakkara D, Abdalla B, Sarkisian C, Goldwater D (2019) The frailty risk score predicts length of stay and need for rehospitalization after kidney transplantation in a retrospective cohort: a pilot study. *Pilot and Feasibility Studies* 5(1):1–9.

Schucht J, Davis EG, Jones CM, Cannon RM (2020) Incidence of and risk factors for multiple readmissions after kidney transplantation. *The American Surgeon* 86(2):116–120.

See KE, Morrison EW, Rothman NB, Soll JB (2011) The detrimental effects of power on confidence, advice taking, and accuracy. *Organizational Behavior and Human Decision Processes* 116(2):272–285.

Sendak M, Gao M, Nichols M, Lin A, Balu S (2019) Machine learning in health care: a critical appraisal of challenges and opportunities. *EGEMs* 7(1).

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.

Shankar N, Marotta P, Wall W, AlBasheer M, Hernandez-Alejandro R, Chandok N (2011) Defining readmission risk factors for liver transplantation recipients. *Gastroenterology & Hepatology* 7(9):585.

Shen J, Zhang CJ, Jiang B, Chen J, Song J, Liu Z, He Z, Wong SY, Fang PH, Ming WK, et al. (2019) Artificial intelligence versus clinicians in disease diagnosis: systematic review. *JMIR Medical Informatics* 7(3):e10010.

Silver D, Hubert T, Schrittwieser J, Antonoglou I, Lai M, Guez A, Lanctot M, Sifre L, Kumaran D, Graepel T, et al. (2018) A general reinforcement learning algorithm that masters Chess, Shogi, and Go through self-play. *Science* 362(6419):1140–1144.

Skjong R, Wentworth BH (2001) Expert judgment and risk perception. *ISOPE International Ocean and Polar Engineering Conference*, ISOPE–I (ISOPE).

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.

Tavares MG, Cristelli MP, Ivani de Paula M, Viana L, Felipe CR, Proença H, Aguiar W, Wagner Santos D, Tedesco-Silva Junior H, Medina Pestana JO (2019) Early hospital readmission after kidney transplantation under a public health care system. *Clinical Transplantation* 33(3):e13467.

Tonekaboni S, Joshi S, McCradden MD, Goldenberg A (2019) What clinicians want: contextualizing explainable machine learning for clinical end use. *Machine Learning for Healthcare Conference*, 359–380 (PMLR).

Tucker LR (1964) A suggested alternative formulation in the developments by hursch, hammond, and hursch, and by hammond, hursch, and todd. *Psychological review* 71(6):528.

Wang Q, Huang Y, Jasin S, Singh PV (2022) Algorithmic transparency with strategic users. *Management Science (forthcoming)* .

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* .

Xin D, Ma L, Liu J, Macke S, Song S, Parameswaran A (2018) Accelerating human-in-the-loop machine learning: Challenges and opportunities. *Proceedings of the Second Workshop on Data Management for End-to-End Machine Learning*, 1–4.

Yala A, Lehman C, Schuster T, Portnoi T, Barzilay R (2019) A deep learning mammography-based model for improved breast cancer risk prediction. *Radiology* 292(1):60–66.

Yaniv I (2004) Receiving other people's advice: Influence and benefit. *Organizational Behavior and Human Decision Processes* 93(1):1–13.

Yataco M, Cowell A, Waseem D, Keaveny AP, Taner CB, Patel T (2016) Predictors and impacts of hospital readmissions following liver transplantation. *Annals of Hepatology* 15(3):356–362.

Yin M, Wortman Vaughan J, Wallach H (2019) Understanding the effect of accuracy on trust in machine learning models. *Proceedings of the 2019 Chi Conference on Human Factors in Computing Systems*, 1–12.

Zeidan JH, Levi DM, Pierce R, Russo MW (2018) Strategies that reduce 90-day readmissions and inpatient costs after liver transplantation. *Liver Transplantation* 24(11):1561–1569.

Zimmerman S, Carder P, Schwartz L, Silbersack J, Temkin-Greener H, Thomas KS, Ward K, Jenkens R, Jensen L, Johnson AC, et al. (2022) The imperative to reimagine assisted living. *Journal of the American Medical Directors Association* 23(2):225–234.

Electronic Companion

EC.1. 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.1 **Summary of AUC performance of ML algorithms considered on the testing set.**

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

EC.2. Human or Algorithm?

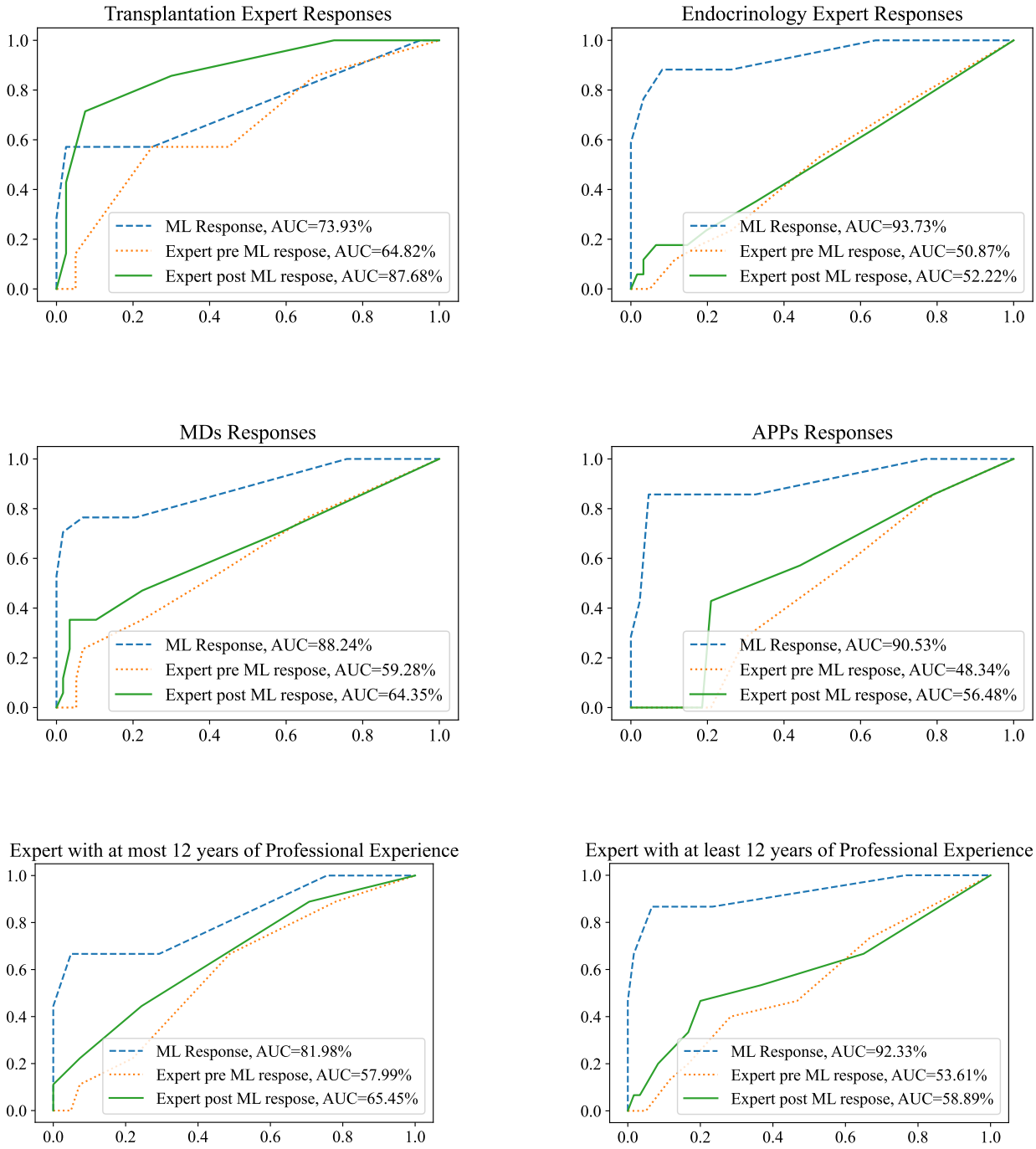


Figure EC.1 Survey Responses ROC Curves.

EC.3. 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.3 **Fleiss' κ measure of inter-rater agreement between human experts on patient clinical characteristics. Only values for features with $\kappa > 0$ are reported.**

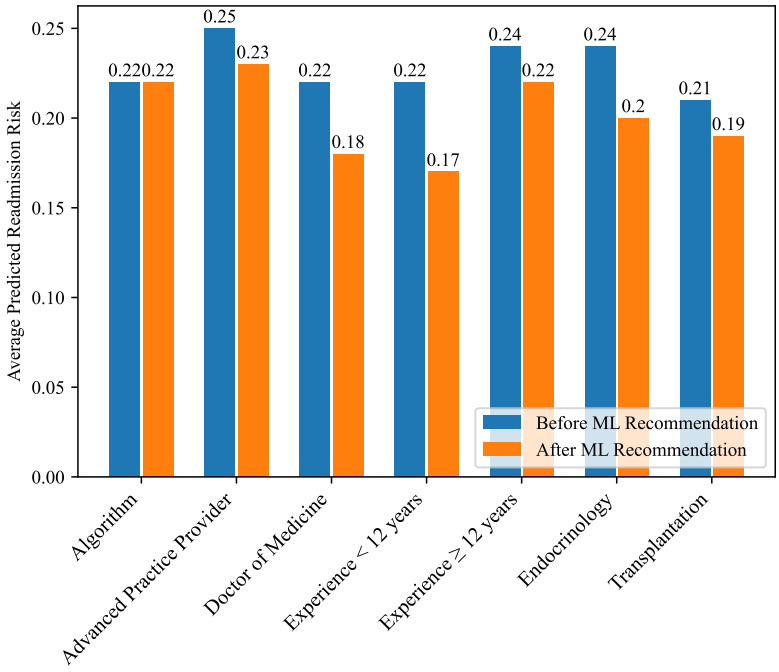


Figure EC.2 **Human experts' risk perception as a function of provider heterogeneity.**

EC.4. Human Expert Risk Perception: A ML Models Comparison

In this Section, we summarize the training and evaluation process that we followed, along with the associated results, to select the algorithm of the human risk perception model. Our dataset comprises all the collected survey responses $(\mathbf{X}_S, \mathbf{Z}_S, \mathbf{y}_S)$ along with the respective patient and human expert features. To train the ML models, we split the data into a training set (75%) and a testing set (25%), stratifying the partition with respect to the outcome of interest $\mathbf{y}_S$. We compare the performance of linear regression with regularization (to avoid overfitting), classification trees (CART), random forests, gradient boosted trees (XGBoost), and support vector machines (SVM) (Hastie et al. 2009, Breiman et al. 2017, Breiman 2001, Chen and Guestrin 2016, Cortes and Vapnik 1995). 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.

Linear regression achieves the best out-of-sample performance compared to the other ML methods considered. For this reason, we select this algorithm for the creation of all human risk perception models presented in the main body of the manuscript.

EC.5. A Bayesian Approach to Measure Human Risk Perception

We conduct a supplemental analysis for the expert risk perception models to compare the frequentist regression with the Bayesian regression approach. We train the models $g'(\mathbf{X},\mathbf{Z}) = \hat{\mathbf{r}}$, $h'(\mathbf{X},\mathbf{Z},f(\mathbf{X})) = \hat{\mathbf{w}}$, and $h''(\mathbf{X},\mathbf{Z},f(\mathbf{X})) = \hat{\mathbf{w}}$. The models for $g'(\mathbf{X},\mathbf{Z})$ and $h'(\mathbf{X},\mathbf{Z},f(\mathbf{X}))$ were both initiated with a randomised prior using a Normal distribution for each independent covariate. In contrast, the $h''(\mathbf{X},\mathbf{Z},f(\mathbf{X}))$ model used as prior the $g'(\mathbf{X},\mathbf{Z})$. Given that questions Q1 and Q5 where provided to the respondents sequentially, our goal with the latter model is to test whether the behaviour of the participants during the first round of risk estimation could more effectively inform the model for capturing their behaviour in the last survey question. Overall, we did not find significant differences between the frequentist and Bayesian models (see Table EC.5). Moreover, the $h'(\mathbf{X},\mathbf{Z},f(\mathbf{X})) = \hat{\mathbf{w}}$ model is marginally more accurate compared to $h''(\mathbf{X},\mathbf{Z},f(\mathbf{X})) = \hat{\mathbf{w}}$. The coefficients of the independent variables in the frequentist approach do not differ significantly to the mean value of the coefficients derived by the Bayesian approach.

Regression Model	$g'(\mathbf{X},\mathbf{Z}) = \hat{\mathbf{r}}$		$h'(\mathbf{X},\mathbf{Z},f(\mathbf{X})) = \hat{\mathbf{w}}$		$h''(\mathbf{X},\mathbf{Z},f(\mathbf{X})) = \hat{\mathbf{w}}$	
Independent Variable	μ	σ	μ	σ	μ	σ
Constant	-0.6095	0.1595	-0.3127	0.1722	-0.4824	0.1028
Patient Information					0.0029	0.0008
Recipient Age at Admission	0.0029	0.0011	0.0028	0.0012	0.0052	0.0020
Recipient BMI	0.0082	0.0029	0.0022	0.0031	0.0055	0.0042
Creatinine Value at Discharge	0.0069	0.0060	0.0043	0.0062	0.0012	0.0003
Average BG Value in Last 24 hrs	0.0007	0.0004	0.0007	0.0005	0.0451	0.0299
History of diabetes 2 mellitus	0.0127	0.0481	0.0151	0.0510	0.0245	0.0102
HbA1c at admission	0.0278	0.0146	0.0237	0.0152	-0.0749	0.0210
Human Expert Information						
Role: MD	-0.0663	0.0317	-0.0711	0.0328	-0.0749	0.0210
Years of Professional Experience	0.0041	0.0016	0.0028	0.0018	0.0035	0.0012
ML Recommendation			0.2536	0.0852	0.2426	0.0805
Model Evaluation						
MAE	0.11		0.1051		0.1081	
Brier Score	0.0204		0.0220		0.0221	

Table EC.5 Output summary of the Bayesian linear regression models capturing the expert risk perception. We report the resulting mean and standard deviation values of the coefficients only for the reduced models.

EC.6. Algorithm or Centaur?

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.

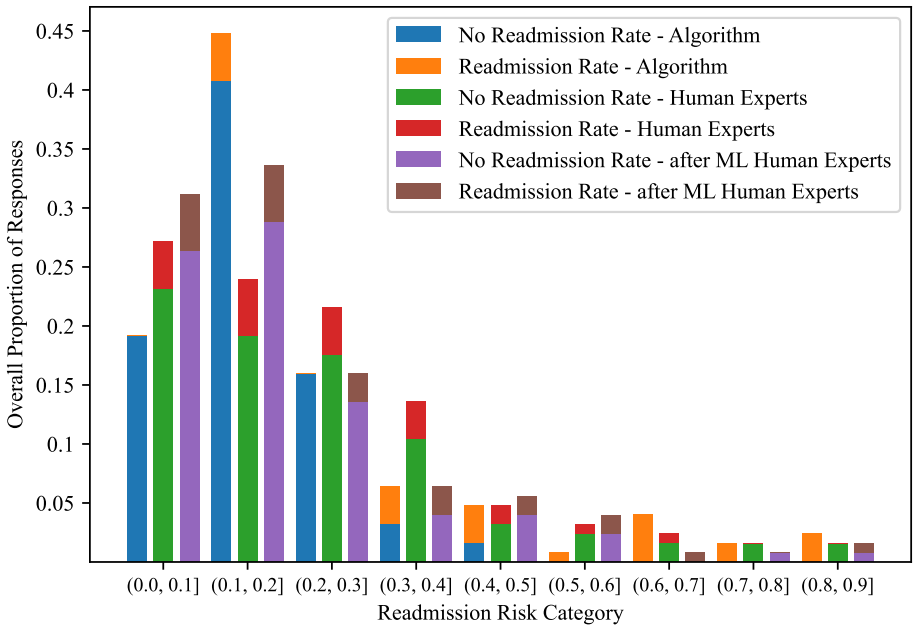


Figure EC.3 Illustration of the overall proportion of algorithm and physician responses in each risk category.

Patient Characteristic	p-value
Kidney	
Creatinine value at discharge	$p \leq 0.0001$
Recipient Age	$p \leq 0.0001$
History of diabetes 2 mellitus	$p \leq 0.0001$
Presence of delayed graft function	$p \leq 0.0001$
Insulin regimen	$p \leq 0.0001$
Average BG Value	$p \leq 0.001$
Recipient BMI	$p \leq 0.001$
Time on dialysis prior to transplant	$p \leq 0.05$
HbA1c at admission	$p \leq 0.05$
Recipient race	$p > 0.05$
Liver	
History of diabetes 2 mellitus	$p \leq 0.001$
Average BG Value	$p \leq 0.001$
Insulin regimen	$p \leq 0.001$
Donor BMI	$p \leq 0.01$
Presence of hyperglycemia during admission	$p \leq 0.05$
Creatinine value at discharge	$p \leq 0.05$
Functional status at transplant	$p \leq 0.05$
HbA1c at admission	$p > 0.05$
Max BG Value	$p > 0.05$
Donor age	$p > 0.05$
Heart	
HLA mismatch level	$p \leq 0.01$
Organ	$p \leq 0.01$
Average BG Value	$p \leq 0.01$
Functional status at transplant	$p \leq 0.05$
Creatinine value at discharge	$p \leq 0.05$
Functional status at listing	$p \leq 0.05$
Presence of LVAD	$p > 0.05$
Deceased donor	$p > 0.05$
Total Ischemic Time (Hours)	$p > 0.05$
History of diabetes 2 mellitus	$p > 0.05$

Table EC.7 T-test p-values for the ten most frequently selected risk factors for each organ that drove clinical risk perception by survey participants.

EC.7. An Application of the Centaur Generalizable Framework

In this section, we illustrate the application of the proposed generalizable framework to our study, illustrating Steps 8 and 9 presented in Section 7.2. 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 EC.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 estimation for the MDs is higher ($AUC(h(\mathbf{X}_R, \mathbf{Z}_{MD}, q(\mathbf{X}_R, l(\mathbf{X}_R)), \mathbf{y}) = 73.54\%$) 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 $AUC(g(\mathbf{X}_R, \mathbf{Z}_{APP}), \mathbf{y}) = 50.76\%$.

Model	Description	$[\mathbf{X}_R]$	$[\mathbf{X}_R, \mathbf{Z}_{MD}]$	$[\mathbf{X}_R, \mathbf{Z}_{APP}]$
$f(\mathbf{X})$	Vanilla ML model	81.61%	N/A	N/A
$q(\mathbf{X}, l(\mathbf{X}))$	Centaur model (no provider heterogeneity)	83.61%	N/A	N/A
$g(\mathbf{X}, \mathbf{Z})$	Expert risk perception model	N/A	55.42%	50.76%
$h(\mathbf{X}, \mathbf{Z}, q(\mathbf{X}, l(\mathbf{X})))$	Centaur-enhanced expert risk perception model	N/A	73.54%	71.83%

Table EC.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.

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 by the capability of the method (eg., linear regression) to capture their choice model. 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 eventual implementation in practice.