HARVARD Kennedy School
JOHN F. KENNEDY SCHOOL OF GOVERNMENT

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

Faculty Research Working Paper Series

Agni Orfanoudaki
Oxford University

Soroush Saghafian
Harvard Kennedy School

Karen Song
Harvard University

Harini A. Chakkera
Mayo Clinic

Curtiss B. Cook
Mayo Clinic

December 2022

RWP22-027

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 a growing amount of 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 an online survey of medical providers and data from a leading U.S. hospital, we develop a ML algorithm and compare its performance with that of medical experts in the task of predicting 30-day readmissions after solid-organ 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. We show that it is able to outperform human experts and the best ML algorithm by systematically enhancing algorithmic performance with human-based intuition. Our results suggest that implementing the centaur model could reduce the average patient readmission rate by 26.4%, yielding up to a $770k reduction in annual expenditure at our partner hospital and up to $67 million savings in overall U.S. healthcare expenditures.*

Key words: Machine Learning, Transplantation, Health care, 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 the *average* human (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 algorithms (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 regulatory 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 (e.g., a decision tree), physicians may disregard it, especially if it involves decisions that contradict medical intuition. These observations raise the question of whether ML modelers need to directly incorporate clinical thinking 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-machine "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)). In it, humans and machines could collaborate

together on the same team. In describing the results, the chess idol Garry Kasparov said 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).

A different but related stream of literature is known as human-in-the-loop, where researchers have proposed various ways to improve the performance of ML algorithms using human input (Xin et al. 2018). 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 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 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 address the following research questions:

1. (Human or Algorithm): *Are humans or algorithms more accurate in predicting 30-day readmissions of solid-organ transplant patients?*

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): *Does combining human experts and ML algorithms improve the performance of ML algorithms?*

5. (Impact on Patient Care): *How would the centaur model (combining human experts and ML algorithms) ultimately affect patient care?*

We contribute to 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 the same information as the algorithm).

• Our study confirms the hypothesis that physician decisions are driven by different clinical features 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 algorithm 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.

• We find that under the centaur model, where human experts interact with the ML algorithm, the human experts' perception of risk improves. However, 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. This suggests the following insight: a little algorithm aversion might be enough to make the centaur model's performance inferior to that of the algorithm.

• We show that when human intuition is systematically incorporated in the ML algorithm, thereby eliminating any impact of algorithm aversion, the performance of the ML algorithm improves. Specifically, the AUC of the ML algorithm improves by 2.46% when it is fed with insights from human experts. The latter finding suggests the following: the centaur model can outperform both the algorithm and the human experts if the impact of algorithm aversion is successfully removed.

• We estimate the potential value that a 30-day readmission predictive tool could bring to the clinical practice by recording the counterfactual decisions that physicians would make at the time of discharge should they have full visibility of the patient's real risk of readmission. We identify that better and closer monitoring of BG values, ensuring caregiver support at home, providing patient education regarding BG control, and extended length of stay are the most common actions that physicians would recommend in order to avoid potential readmission when the algorithm flags a patient as high risk. Finally, we find that by doing so, physicians might be able to reduce

readmission rates by 26.4% compared to the current practice. Given that each readmission among transplanted patients is estimated to cost about $27,000 (Weiss and Jiang 2021), this reduction in readmissions can translate up to about $770,000 less expenditure per year at our partner hospital. This also implies that implementing the centaur model nationally could yield a reduction of about $67 million per year in overall healthcare expenditures in the U.S.

The remainder of the paper is organized as follows. Section 2 provides a summary of the literature relevant to our research questions outlined above. Section 3 introduces our ML algorithm 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 the proposed 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 human-in-the-loop approach that allows us to augment the ML algorithm with guidance from expert clinicians. In Section 8, we perform a counterfactual analysis guided by the survey responses to gauge the impact the centaur model could bring to practice. 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.

There is an increasing number of studies suggesting 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 Supreme Court of the United States. 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 at 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)).

Yet, the degree and the factors that affect 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 (2021) promoted a "two-way personalization" model, whereby incorporating preferences of physicians into a causal inference algorithm, recommended treatment plans are personalized both 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. We propose the integration of expert advice into the ML system in the form of an exogenous predictive model that is trained on historical data of human judgment.

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, an early readmissions risk score after any solid organ 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 2021). However, the role of these factors in the context of early hospital readmissions has not been studied yet. We aim to explore the relationship between inpatient glucose control and hospital readmissions among all types of solid organ transplantation, incorporating patient factors that are either specific or shared among the three organs.

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 operational records of the hospital. We supplemented this data with donor and organ-specific information from the United Network for Organ Sharing (UNOS) registry. Finally, as discussed 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

Most of the patients in our data set had kidney transplantation (67.5%) 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. Overall, 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).

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 (126.0-171.0)	All	MELD Score	18.0 (12.0-25.0)	Liver
BG Average Value Last 24hrs	143.0 (126.0-173.0)	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.

3.2. The Machine Learning Algorithm

We train 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. 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).

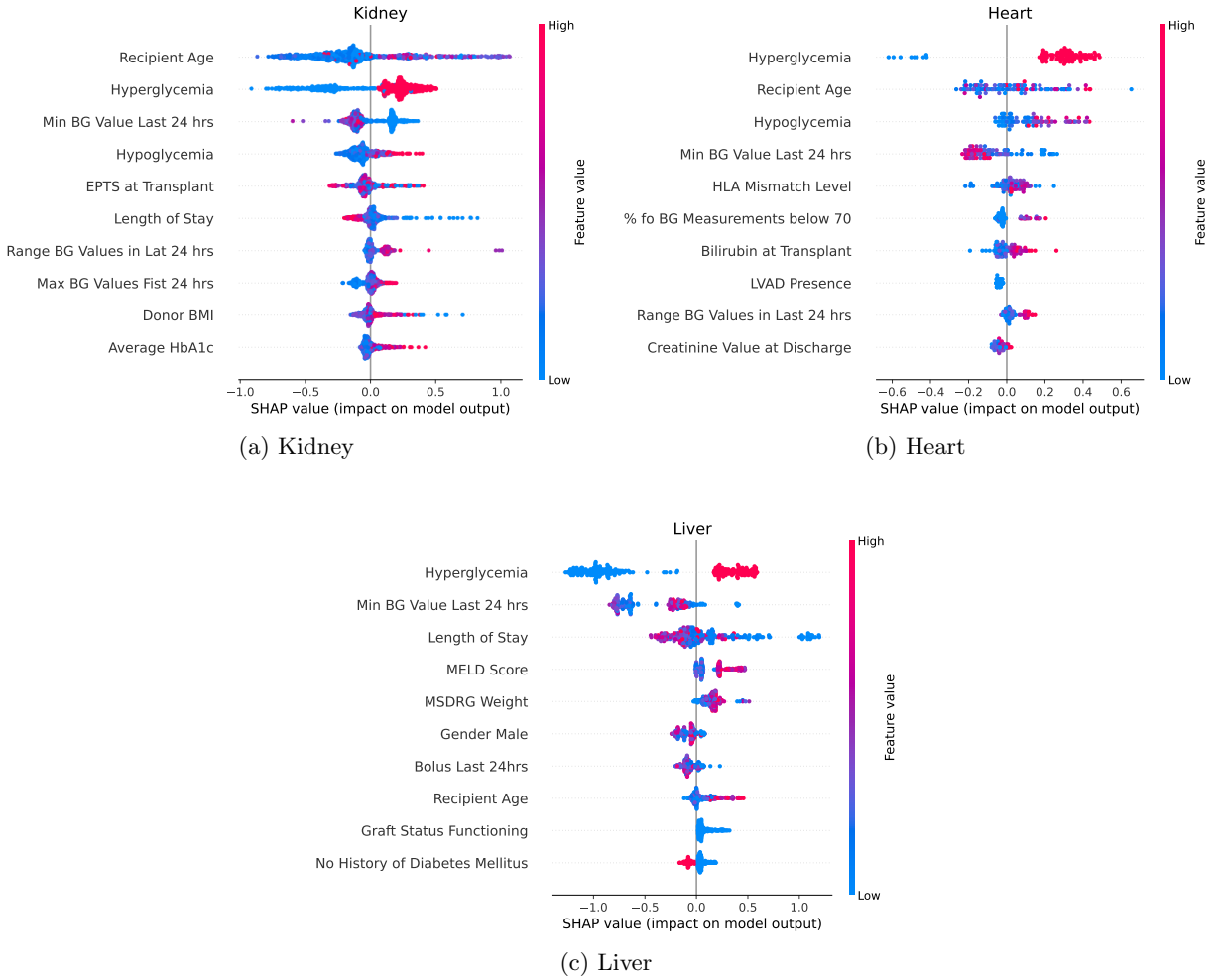
We compute the average out-of-sample value and standard deviation of the AUC. Our computational results demonstrated that the XGBoost algorithm achieves superior performance compared to the other methods considered (see Table EC.1). The XGBoost models achieve a mean 84.0% AUC on the testing set with 0.05% standard deviation. Our analysis suggests that the models' AUC is higher (86.8%) for patients with a history of diabetes mellitus. The downstream performance of the 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) is the main reason behind the lower AUC value for these patients. Nevertheless, our ML algorithm achieves better 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. Clinical Insights

We use the SHapley Additive exPlanations (SHAP) framework to derive clinical insights from the model predictions (Lundberg and Lee 2017, Lundberg et al. 2020). Our goal is to identify the main independent variables that can predict early hospital readmission per organ type. In addition, we use this tool to test our hypothesis of whether metabolic factors are associated with worse patient outcomes. In Section 6, we compare these findings with physician responses, questioning whether ML and human intuition are aligned.

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.



(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.

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 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.

The ML algorithm also uncovers the role of other risk factors. 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 MELD score, MSDRG weight, recipient gender, and functioning status of the graft (Yataco et al. 2016). Last but not least, in the case of heart transplantations, the model identifies the HLA mismatch level, the values of bilirubin and creatinine, and the presence of LVAD as highly predictive features (Kim and Kim 2020).

4. Survey Design

To answer our research questions, presented in Section 1, 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.

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? 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

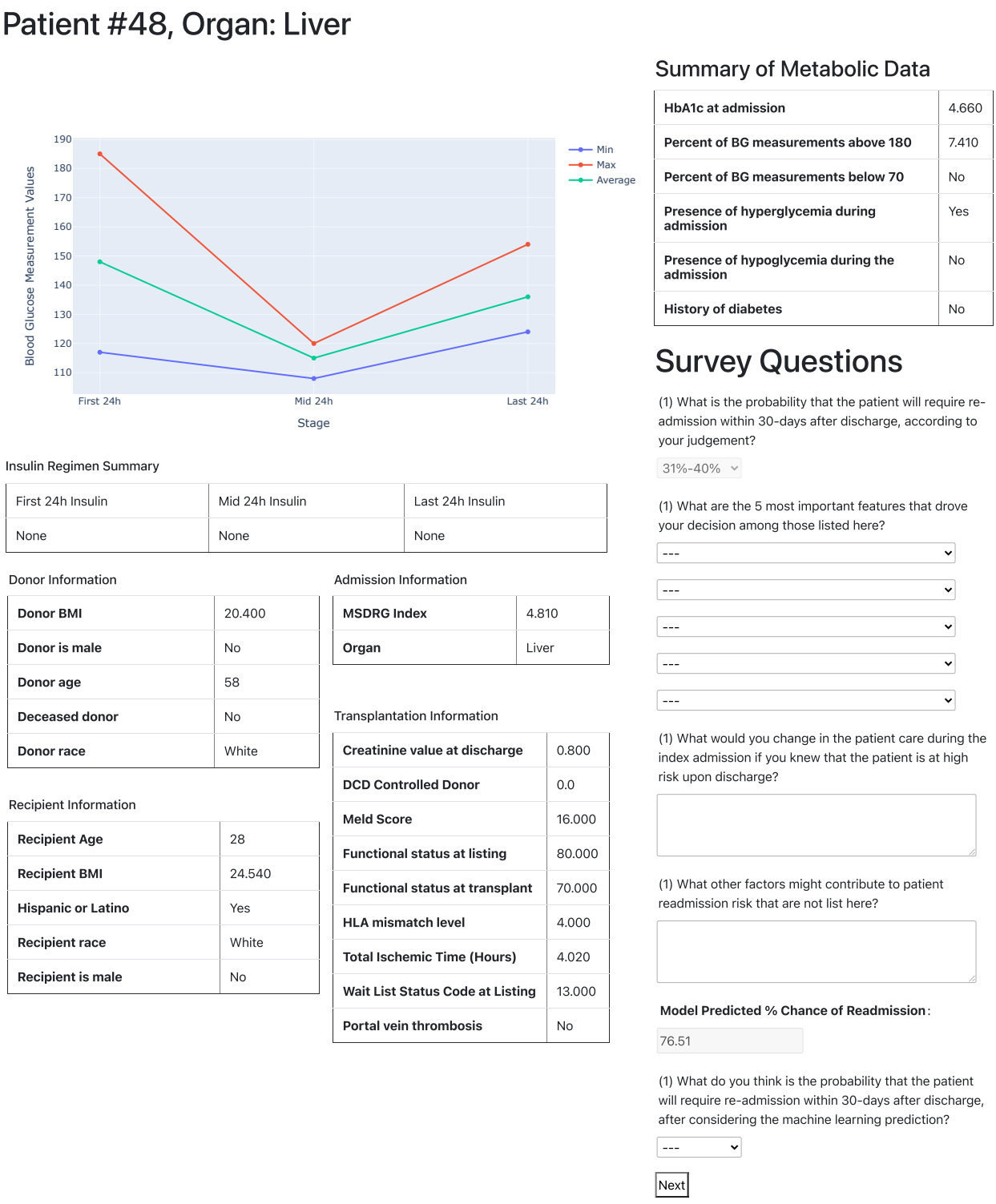


Figure 2 Illustration of the survey tool interface for an example liver patient.

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 at Harvard University. The study was also approved by both Harvard's and Mayo Clinic's Institutional Review Boards. 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 out-of-sample accuracy of the proposed risk score. 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.

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 after finalizing their initial estimation to ensure non-biased responses. 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

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.

5.1). 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 find 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. In answering the first survey question (Q1), participants were asked to provide their risk estimation using intervals with 10% increments (e.g., $[0\%, 10\%), [10\%, 20\%)$, etc.). 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. 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, 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 are statistically significant with p-values< 0.001.

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

biased 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% (14.89%) and 43.04% (14.67%) respectively). This is reflected in the relative AUC improvement of the former (35.26%) but less so in the performance of the latter (8.55%). In fact, APPs achieve double relative improvement (16.84%) compared to MDs with substantially lower WoA (26.36% (14.0%)). We do not identify significant WoA differences between medical practitioners with less or more 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.

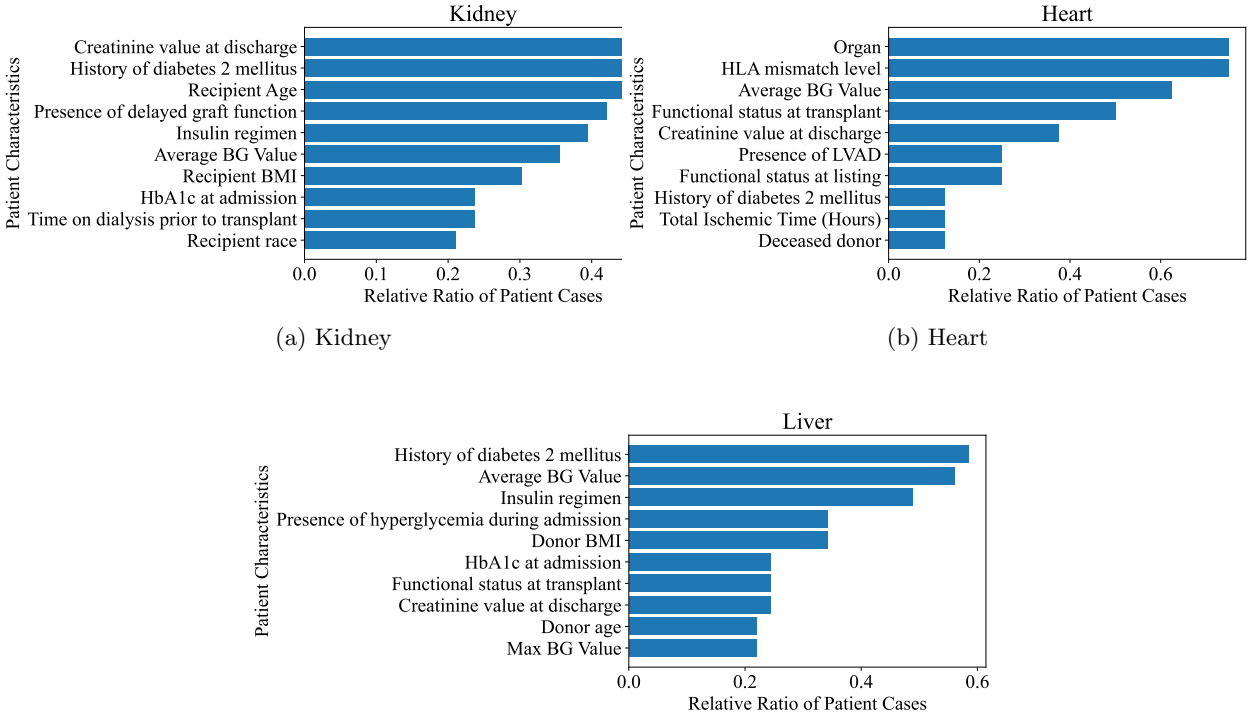
Put together, our results show that, when provided with exactly the same information, human experts are less accurate compared to a ML algorithm. However, our analysis reveals that providing the algorithm's estimation as an input to clinicians at the time of the decision can positively influence their 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 that participants who are more willing to consider the data-driven model are associated with better results.

6. Reasoning and Risk Perception

In this section, we aim to focus on the third research question; 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 take a human-centered perspective; reporting 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



(a) Kidney

(b) Heart

(c) Liver

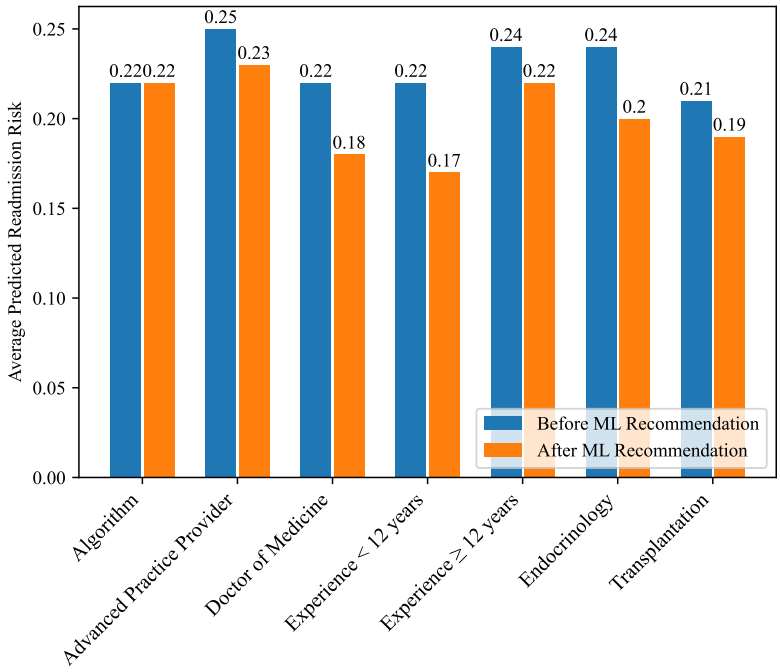
Figure 3 Relative frequency of reported drivers of 30-day readmission risk perception based on the survey responses. Acronyms are defined at the notes of Table 1.

For each case, survey participants were asked to report five main patient characteristics that drove their risk estimation. Our goal was to uncover the perceived drivers of readmission risk from medical practitioners and compare 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.4. 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

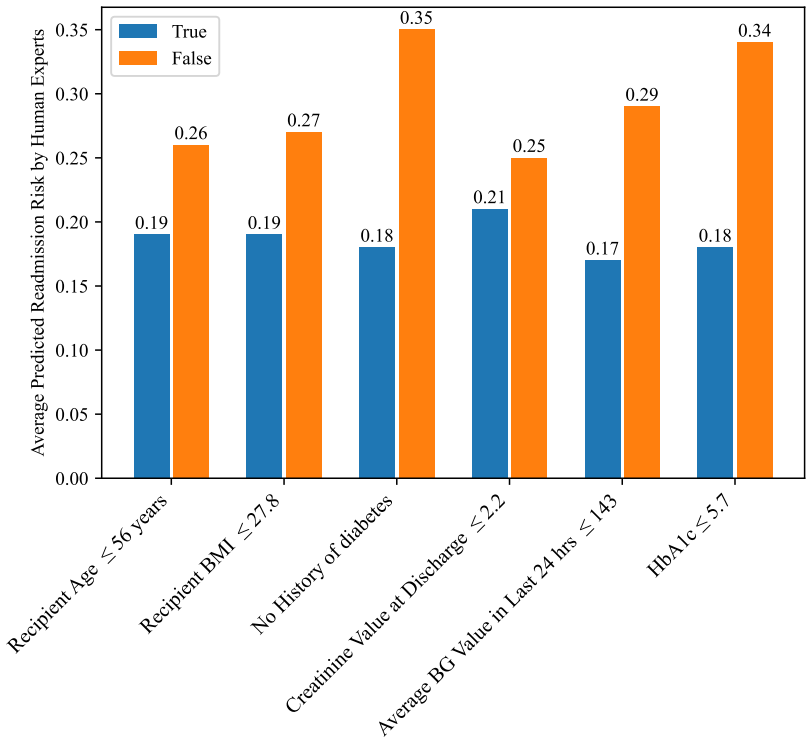
also included in 20% to 30% of the responses. In the case of heart transplantations, the type of organ and the HLA mismatch level were selected in more than 70% of patients. Similarly to kidney patients, the average BG value during the stay and creatinine value at discharge were reported in at least 40% of the survey submissions. The functional status at transplant and listing, the total ischemic time, and the presence of LVAD 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 medical experts. The proposed 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 and history of diabetes as two of the most important determinants of readmission risk. The judgment of medical providers was also driven to a higher degree by organ-specific variables, including the presence of LVAD and the organ's functional status at transplant and listing.

Our findings provide evidence that humans do not tend to place as much emphasis on metrics that capture fluctuation and variability but rather 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 to measure the inter-rater agreement between participants for each individual patient, we observed a high degree of consensus ($k > 0.2$) for only a small subset of variables (see in Appendix EC.3) (Fleiss 1971). For example, in more than 50% of the patient cases reviewed, the presence or absence of history of diabetes was one of the five selected variables by the participants. Nevertheless, there was very 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.



(a) Human Experts Characteristics



(b) Clinical Risk Factors

Figure 4 Human experts' risk perception as a function of provider and patient heterogeneity.

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 when needed. 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, to complement the answer to our third research question, we distinguish between two types of risk misperception—overestimation and underestimation— and shed light on both provider and patient characteristics that can yield one type versus the other.

First, in Figure 4, we present the average predicted risk for different provider and patient population subgroups. The overall predicted risk from the ML algorithm was slightly lower (22.36%) compared to the human experts' estimations in the first round of responses (23.16%). In the second round of responses, the introduction of the ML suggestion led to a decrease in the average predicted risk for the sample population (20.04%). The survey responses revealed that overall APPs, endocrinology specialties, and providers with at least 12 years of experience are more conservative than the ML model as the average predicted patient risk was at least 24%. While the average estimated risk reduced after the ML recommendations across all human expert categories, MDs, endocrinology specialists, and providers with less than 12 years of experience were most significantly impacted. In the latter case, the average predicted patient risk reduced from 22% to 17%, while in the case of MDs, it dropped by 4%. Our analysis corroborates the hypothesis that provider heterogeneity affects the propensity of human experts to overestimate and underestimate patient risk.

Next, we focus on some of the clinical characteristics that physicians highlighted as the primary drivers of their risk perception (see Section 6.1). First, Figure 4b validates the physicians' own responses as we see that the highest average readmission rates are reported for the clinical features that experts independently identified. Our analysis shows that the baseline risk for organ recipients of lower (higher) age and BMI was 19% (26-27%). Patients with no history of diabetes and high hemoglobin A1c values were associated with the highest perceived risk, ranging between 34% to 35%. In addition, we see that high values in laboratory test results like creatinine levels at the time of discharge and BG measurements during the last 24 hours of the patient stay are associated with a significantly higher perception of risk by human experts.

Although these results highlight the overall perception of the experts compared to the algorithm, they do not specify whether humans overestimate or underestimate risk as a function of true patient outcomes. In Table 3, we focus on all participant responses and present the proportion

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.

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 expectations regarding the patient trajectory and predicted 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. It also highlights that although providers, on average, predicted higher risk for the patient population compared to the algorithm, for the majority of patients who required readmission, their perception of risk was less conservative than the ML algorithm. Therefore, we observe that human experts have an edge over the algorithm in a subset of patients. Thus, there could be potentially valuable human insights that could be transferred to the ML model.

These associations prompt us to investigate whether clinical intuition could be captured in a more systematic form. For this reason, in the next section, we test if a data-driven model can accurately predict the human experts' risk perception.

6.3. Capturing Human Intuition: A Regression Model

We develop a linear regression model capable of predicting and inferring the experts' estimation of the risk of readmission. We use as independent variables the patient information described in

Independent Variable	**Regression Coefficient**	**P-value**	**2.5% Q**	**97.5% Q**
Constant	-0.6169	<0.001	-0.932	-0.302
Patient Case Information				
Recipient Age at Admission	0.0029	0.012	0.001	0.005
Recipient BMI	0.0083	0.004	0.003	0.014
Creatinine Value at Discharge	0.0071	0.023	0.002	0.019
Average BG Value in Last 24 hrs	0.0017	<0.001	0.001	0.003
History of diabetes 2 mellitus	0.0124	0.008	0.011	0.03
HbA1c at admission	0.0285	0.004	0.015	0.057
Human Expert Information				
Role: MD	-0.0673	0.032	-0.129	-0.006
Years of Professional Experience	0.0041	0.013	0.001	0.007

Table 4 Output summary of the linear regression model. We report the resulting coefficients only for the reduced model with statistically significant t-tests values.

Section 3.1. The providers' responses to the first survey question (Q1) are used as the dependent variable. Our model aims to predict the continuous risk estimation provided by the experts that ranges from zero to one. Our goal is to test whether a simple linear regression model can capture medical reasoning and summarize humans' risk estimations. In addition to the patient characteristics, we include the expert's specialty information in the set of independent variables. Given that some patients were reviewed by two experts of different specialties, we developed a model that accounts for physician heterogeneity. Thus, we consider each survey response as a distinct observation.

We use ordinary least squares regression to predict the continuous risk score provided by the survey participants (Hastie et al. 2009). We removed from the model all the independent variables with insignificant t-tests. For the 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). In the final model $R^2 = 0.85$ and the adjusted $R^2 = 0.76$. 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, and the 95% confidence intervals, 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. Overall, we provide further evidence of provider heterogeneity in risk perception and attribution.

Put together, our results validate the hypothesis that the judgment of human experts can be captured by a linear model with very high precision. This finding is in line with past scientific studies that have claimed that superior ML discrimination power in complex tasks in the field of medicine can be attributed to the ability of such models to capture non-linear relationships in the data (Orfanoudaki et al. 2020, 2022). The presence of a linear mental model of risk for clinical practitioners can be explained by the tools the medical community has employed over the past 70 years to establish risk associations between patient characteristics and potential adverse events. The vast majority of clinical studies have employed linear models, such as ordinary least squares, logistic regression, and the Cox proportional hazards model, to identify the role that potential risk factors can play in disease diagnosis and evolution. As a result, physicians and APPs might have been trained to delineate such associations and think in the same "linear" manner. Our work demonstrates that the deployment and integration of ML tools can complement the currently established perceptions of risk, personalizing risk attribution and identifying non-linear interactions.

Thus, we showed that clinical intuition could be captured in the form of an organ-specific linear regression model. In the next section, we illustrate how feeding such a regression model to the original ML algorithm can help us move towards a human-algorithm "centaur" model that is 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 findings from the survey, we investigate whether we can improve the ML algorithm's performance using structured human intuition. This could allow us to recommend a human-algorithm "centaur" model that can outperform both the human experts and the ML algorithm.

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. This negatively impacted the model's performance since the R^2 dropped to 0.80 and the adjusted R^2 to 0.72. The updated regression coefficients can be found in Table EC.3. To incorporate human intuition, we apply the resulting linear model to the entire sample population and use its output as an additional independent variable for the downstream ML algorithm prediction. The expected value of the resulting feature is 23.0% with standard deviation of 18.0%. By attaching this value to the patient vector across all observations, we augment the feature space provided by the hospital's health records with a composite variable that summarizes the experts' perception of risk based on the same set of features.

Subsequently, we re-train our XGBoost algorithm following the same process outlined in Section 3. We split the data into the same partitions to derive a fair comparison of predictive performance. The average out-of-sample AUC of the new classifier across all random data partitions is 86.46%, which is a 2.46% improvement compared to the original ML algorithm. The performance improves by 2.46% compared to the original model which did not include the experts' choice model. By analyzing the performance across the three types of organs, we observe that the greatest benefit is manifested in the case of kidney transplantations. The AUC of the human-algorithm model increased to 85.1% compared to 77.0%. This finding provides concrete evidence that incorporating expert insights in the form of a model can substantially improve algorithmic performance.

We do not find substantial changes in the case of liver and heart patients (97.6% and 66.8% respectively). The AUC of the original XGBoost model for liver transplants is as high as 97.5%, and thus, sustaining a further improvement remained a more challenging task compared to the case of kidney transplantations. Conversely, the latter formed the majority of patient cases that were reviewed by the survey participants, and the baseline AUC performance of the ML algorithm was substantially lower (77.0%). As a result, human intuition is expected to yield a greater benefit for this subgroup of patients. Unfortunately, the sample size for heart patients both in the case of the survey study as well as in the hospital data set is very small—heart transplantation is a relatively rare operation. This limitation can potentially explain why our approach does not substantially differentiate the discrimination performance for this patient subgroup.

This experiment affirms that a centaur model that incorporates human insights and intuition into the algorithm development and validation can improve the downstream performance of the algorithm. Our work introduces a human-in-the-loop approach for developing the centaur model 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 proposed. It demonstrates a scalable and effective way to boost algorithmic performance without constant human supervision. However, 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.

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?

8. The Impact on Patient Care

We now answer the fifth research question we raised in Section 1: what is the impact that the proposed ML model could have on patient care? 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 our findings. 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%). In measuring readmission risk, the degree to which these results affect provider decisions determines the scope and value of such models. In this setting, 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. Even though we do not find significant differences between the endocrinology and transplantation teams, 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 place a lot of emphasis on treatment education to ensure high adherence to post-transplantation and BG therapy. MDs favor 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 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. The Mayo Clinic could integrate the proposed ML model into its electronic health records system and use it to flag patients at high risk at the time of discharge. Our analysis highlights five broader categories of action providers would be keen to follow to avoid potential future readmissions. The hospital system could establish a set of processes to which MDs and APPs could directly refer patients, including (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. Such processes could operationalize ML model estimations, directly tying risk evaluation to decisions that affect the care pathway.

To estimate the potential operational and financial impact that the readmission risk score could bring to the organization, we focus only on two of the actions proposed by the physicians, namely "Improve glycemic control" and "Extend hospital length of stay," which depend on provider decisions only during the index admission. We evaluate the expected reduction in the readmission rate if all high-risk patients were provided additional metabolic treatment that regulated their BG values during the last 24 hours of their index admission at the expense of one additional day in the hospital. This provides a conservative estimate of the overall benefit of the algorithm since (a) it does not consider the impact of the other alternatives mentioned, and (b) it increases the length of stay of all patients under high risk even though some patients' metabolic therapy could be regulated at the time of discharge. To derive our estimate, we assume the following: (i) physicians have

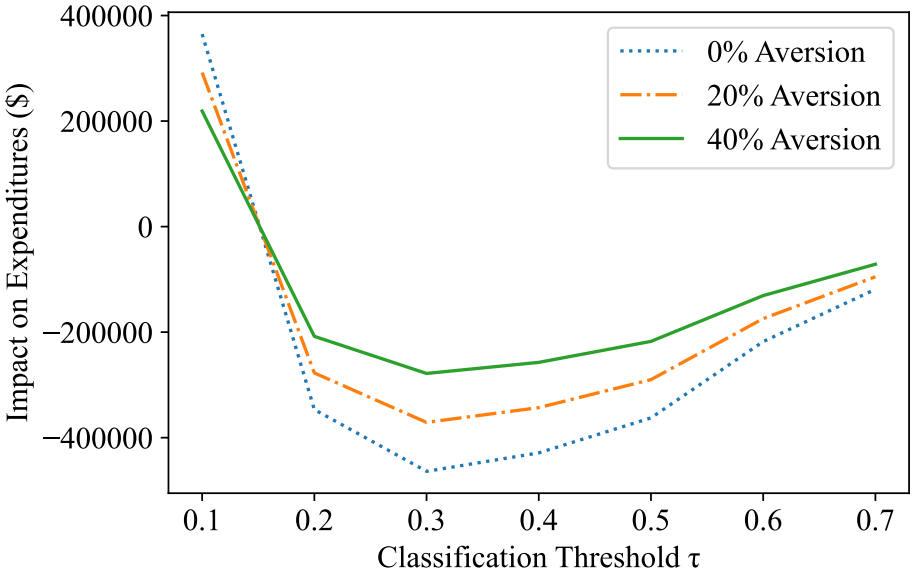
the capability to use basal and bolus insulin to fully regulate the BG levels during the last 24 hours of the stay; (ii) the changes in the metabolic regimen do not impact other clinical features of the patient (e.g., creatinine levels); (iii) there is a certain classification threshold τ such that for any predicted risk levels $> \tau$ medical practitioners provide additional metabolic treatment to regulate BG levels. We relax the latter assumption to estimate the respective impact under different degrees of algorithm aversion.

τ	**Sensitivity**	**Specificity**	**Adjusted RR**	**Subgroup RR**	**Subgroup Adjusted RR**	**Patient %**
0.1	96.05%	17.48%	15.87%	26.66%	17.33%	85.19%
0.2	81.58%	69.58%	15.96%	40.44%	21.04%	40.52 %
0.3	73.68%	86.08%	16.92%	49.28%	22.47%	25.71 %
0.4	59.21%	93.2%	18.28%	56.02%	23.76%	17.14 %
0.5	46.05%	96.76%	19.49%	61.59%	24.58%	11.69 %
0.6	23.68%	98.71%	21.38%	69.16%	26.49%	5.71 %
0.7	10.53%	99.35%	22.55%	76.65%	27.93%	2.6 %

Notes. The Table presents the results for different levels of the classification threshold τ as well as the respective sensitivity and specificity of the algorithm on the testing set. The Adjusted 30-Day Readmission Risk (RR) columns indicate the average RR of the entire patient population after the model's development. The Subgroup columns refer to the average RR with and without the ML intervention for those patients whose predicted score was higher than τ.

Table 6 Summary of the proposed algorithm's impact on the average RR for the entire patient population and the patient subgroup that experiences a change in treatment and extended length of stay.

We summarize the results of our counterfactual analysis in Table 6 and Figure 5. We consider seven different values for the classification threshold τ. Lower τ values improve the model's sensitivity, while higher values favor its specificity. To measure the counterfactual effect, we hypothesize that every patient whose predicted RR is higher than τ will stay one additional day at the hospital and receive BG treatment to regulate their metabolic factors. The last column of Table 6 reports the proportion of patients that meet this criterion. We apply this approach to the testing set of the sample population. We modify the BG related variables that refer to the last 24 hours of the patient stay and set them to the mean value of the overall population as described in Table 1. We get the new predicted RR for those observations using the proposed algorithm and measure the difference. The baseline RR for all patients without the ML model is 23.0%. Table 6 outlines the impact on the mean readmission rate of the entire population and the patient subgroup that received the intervention. Our analysis shows that patients can significantly benefit from the targeted modifications in patient care as the RR can be substantially reduced. For example, in the case of $\tau = 0.3$, the overall RR drops from 23.0% to 16.92%. The algorithm impact becomes even



Notes. The vertical axis reflects the change in cost for the healthcare system. Positive values indicate increases with respect to the baseline, while negative values refer to reductions. The percentage of aversion corresponds to the inverse proportion of providers following the algorithm recommendation. Thus, when aversion is equal to 20%, 80% will experience the extended length of stay and changes in the BG treatment.

Figure 5 Impact of algorithmic deployment on healthcare expenditures.

more evident when we focus only on the subgroup that is labeled as high-risk. For the same setting, the average RR drops from 49.28% to 22.47%.

We highlight that the estimated reductions in 30-day readmissions can have significant financial implications for many hospitals, including the Mayo clinic. Past evidence from the literature suggests that the average readmission cost for index admissions related to complications of transplanted organs or tissue per patient was $27,000 (Weiss and Jiang 2021). In addition, a past study on kidney transplantations found that the average variable cost of an additional day at the hospital during the index admission was $3,422. Thus, successful changes in patient care would result in savings of $23,578 (true positive benefit), while unnecessary alerts for rather healthy patients would incur an additional cost of $3,422 (false positive penalty). Using these numbers, we find that, by implementing our algorithm, the Mayo Clinic unit could reduce its annual expenditures by about $463,778 (assuming $\tau = 0.3$).[1] These remain conservative estimations since we assume that all patients with risk scores above τ will require an additional day at the hospital. If we hypothesize that optimal BG control could be achieved without any extension in the length of stay, the cost

[1] This estimate is based on the assumption that there is no algorithm aversion (i.e., all identified patients receive the intervention). We relax this assumption for two different levels of algorithm aversion to estimate its impact on the potential benefit the algorithm could yield. We find that for the degree of algorithm aversion derived from the survey platform (40%), the hospital could expect savings of approximately $278,266.

savings for Mayo could be up to $770,000. More broadly, if our algorithm is implemented nationally, we find that the overall expenditure related to kidney, liver, and heart transplantation in the U.S. could be reduced by 67 million based on the total annual number of organ transplants in the country (OPTN 2022). Thus, we hope to see a broader implementation of our proposed algorithm across hospitals and medical centers.

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 centaur algorithm to bridge the gap between the two. Contrary to other approaches, a human-algorithm centaur model can involve the human-in-the-loop prior to algorithmic deployment. We show that the centaur is better than independent algorithmic estimations and the human experts' evaluations. This finding is partly driven by the fact that expert intuition can complement machine insights. Even though our analysis was based on an empirical study in the clinical setting, the centaur approach is applicable even beyond the context of healthcare. Future work could focus on validating our findings in other domains where algorithms are called to enhance human experts' decisions, such as legal practice.

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. 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). Survey participants highlighted that there are other confounding factors that are not present in the study, 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 potential confounders.

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.

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.

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) Pricing 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.

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).

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.

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).

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.

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`.

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.

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`.

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.

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.

OPTN (2022) All-time records again set in 2021 for organ transplants, organ donation from deceased donors - OPTN. optn.transplant.hrsa.gov/news/all-time-records-again-set-in-2021-for-organ-transplants-organ-donation-from-deceased-donors/.

Orfanoudaki A, Chesley E, Cadisch C, Stein B, Nouh A, Alberts MJ, Bertsimas D (2020) Machine learning provides evidence that stroke risk is not linear: The non-linear framingham stroke risk score. *PloS one* 15(5):e0232414.

Orfanoudaki A, Dearani JA, Shahian DM, Badhwar V, Fernandez F, Habib R, Bowdish ME, Bertsimas D (2022) Improving quality in cardiothoracic surgery: Exploiting the untapped potential of machine learning. *The Annals of Thoracic Surgery* .

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).

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 (2021) Ambiguous dynamic treatment regimes: A reinforcement learning approach. *arXiv preprint arXiv:2112.04571* .

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.

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).

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.

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).

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.