Development of explainable artificial intelligence based machine learning model for predicting 30-day hospital readmission after renal transplantation

Nasser Alnazari[1,2*], Omar Ibrahim Alanazi[3], Muath Owaidh Alosaimi[3], Ziyad Mohamed Alanazi[3], Ziyad Mohammed Alhajeri[3], Khaled Mohammed Alhussaini[3], Abdulkarim Mekhlif Alanazi[3] and Ahmed Y. Azzam[4]

Nat Sirirutbunkajorn
Radiation oncologist, Ramathibodi hospital
Department of Clinical Epidemiology and Biostatistics, Faculty of Medicine Ramathibodi Hospital, Mahidol university
Nut19012537@gmail.com

Background

- **Hospital re-admission** is a challenge after renal transplantation, reflecting complication from transplant.
- **Standard clinical approaches** to predicting readmission risk **may not fully capture the underlying relationships** precisely for patients.
- Our study addresses this by developing an **explainable artificial intelligence** (XAI) for predicting **30-day hospital readmission risk** following renal transplantation
  - including both of pre-transplant and post-transplant variables to create a **clinically applicable prediction tool.**
  - not only focusing on the predictive accuracy but **also model interpretability**, to ensure that healthcare providers can understand and **trust the factors driving the predictions** from the developed model

Study design

- **Retrospective study from King Abdullah International Medical Research Center (KAIMRC)**
- **Inclusion criteria:**
  - Adult who underwent renal transplantation, living and deceased donor
  - N = 588 patients
- **Data source**
  - EMR
    - Demographic information, clinical variables, laboratory values, and transplant-specific characteristics
- **Label:**
  - Admission withing 30-days after discharge
    - **All hospital encounters, including observation stays and emergency department visits without formal admission**
    - Event number = 523/588 (89%)

Baseli

Characteristic:	Value / Number:
Baseline characteristics:	
Age, Mean (SD)	54.3 (12.6)
Total cohort size	588
Follow-up duration, days	11.2 ± 17.9 [6.0 (1.0–13.0)]
Male	367 (62.4%)
Female	221 (37.6%)
Body Mass Index, kg/m^2	26.2 ± 6.1 [26.4 (21.9–30.3)]
Transplant-Related Characteristics:	
Living donor	500 (85.2%)
Deceased donor	87 (14.8%)
A	218 (37.1%)
B	136 (23.2%)
AB	121 (20.6%)
O	31 (5.3%)
Immunosuppression regimen	563 (95.9%)

Baseli

Pre-transplant Clinical Parameters:	
Systolic blood pressure, mmHg	135.0 ± 22.3 [136.0 (120.0–150.0)]
Diastolic blood pressure, mmHg	76.6 ± 15.1 [77.0 (66.0–87.0)]
HbA1c, %	5.8 ± 1.5 [5.3 (4.9–6.1)]
eGFR, mL/min/1.73 m^2	13.3 ± 18.5 [7.0 (5.0–12.0)]
Diabetes Mellitus	341 (58.3%)
Post-transplant Clinical Parameters:	
Systolic blood pressure, mmHg	135.5 ± 19.4 [137.0 (122.0–149.0)]
Diastolic blood pressure, mmHg	74.5 ± 15.5 [75.0 (63.8–84.0)]
HbA1c, %	6.2 ± 1.6 [5.7 (5.2–7.1)]
eGFR, mL/min/1.73 m^2	19.2 ± 18.5 [13.0 (9.0–21.0)]
Serum creatinine, mg/dL	482.9 ± 275.8 [441.0 (293.0–643.0)]
Outcomes and Complications:	
Length of initial hospital stay, days	4.3 ± 5.0 [3.0 (1.0–6.0)]
Readmission rate within 30 days	2.2 ± 2.6 [1.0 (1.0–2.2)]
Patients requiring readmission	523 (88.9%)
Graft rejection episodes	42/70* (60.0%)

Preprocessing pipeline

Handle missing value with MICE

Five imputations for continuous variables and mode imputation for categorical variables with less than 20% missingness
Exclude variable with >20% missing

Standard scaling (mean = 0, SD = 1) continuous variables
Encode all categorical variable

Values > 3 SD = Outlier -> Winsorization

Application of clinically validated threshold

Preprocessing parameter derived from the training set only for each CV

Feature Selection

What domain knowledge? Keep what feature with VIF > 5?

Domain knowledge

Univariate analysis (p-value < 0.2 threshold)

Handle multicollinearity: Remove variance inflation factor over five

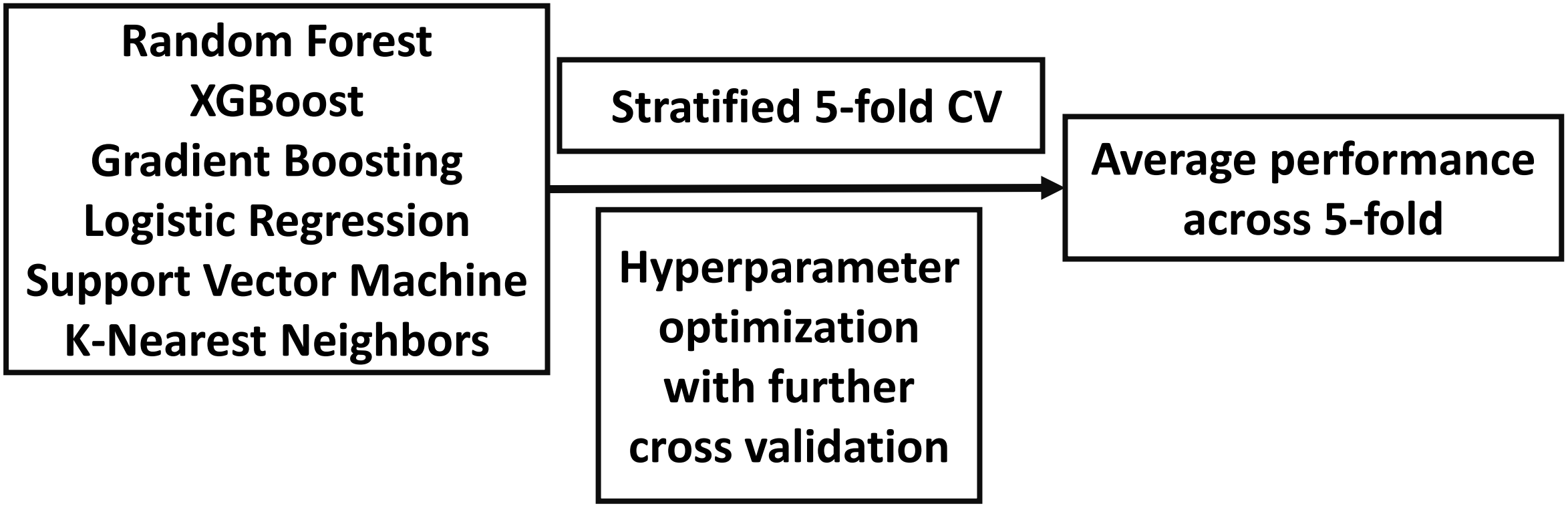
168,596 data points → 15 predictors

Feature Selection

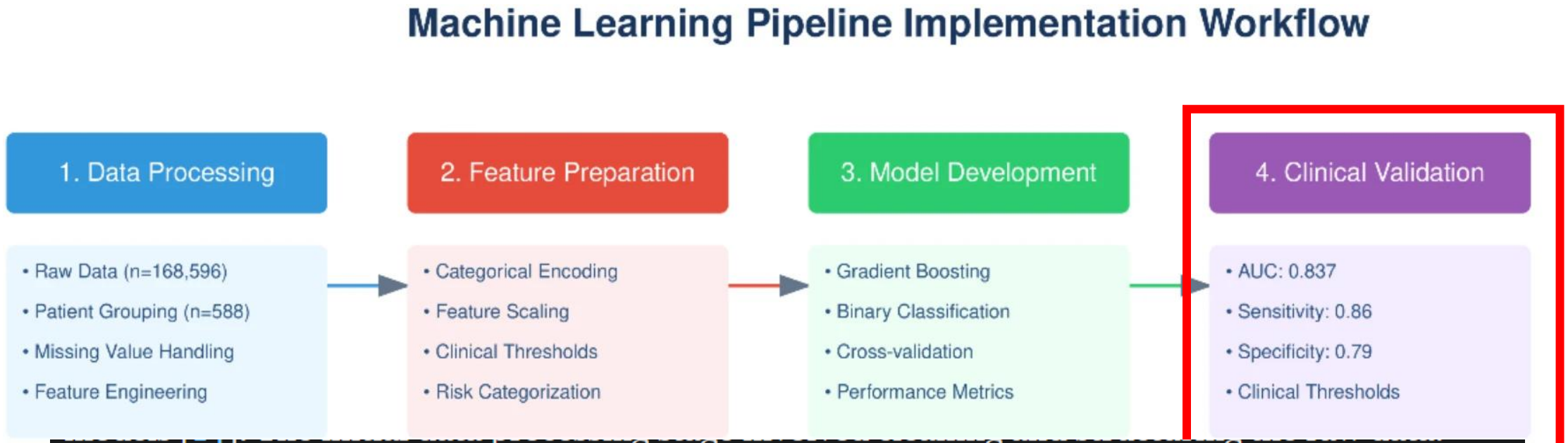
preserve sample size. Feature selection combined clinical domain knowledge with statistical filtering using univariate analysis (p-value < 0.2 threshold) and assessment of multicollinearity (removing features with variance inflation factor over five), ultimately reducing our initial 168,596 data points to 15 finalized predictors.

Feature selection on whole dataset = cheating!

Algorithm selection and training



Algorithm selection and training



The feature preparation phase included categorical variable encoding, feature scaling to standardize numerical variables, and implementation of clinical thresholds based on established medical guidelines. The model development stage employed binary classification approaches with cross-validation methodology. Clinical validation was performed as the final stage to ensure and validate the medical relevance and practical applicability.

Algorithm selection and training

Implementation pipeline architecture and model deployment

The implementation framework was executed through a four-stage pipeline, initiating with the processing of 168,596 raw data points derived from our cohort of 588 patients. The feature preparation phase has included advanced categorical encoding techniques, standardized feature scaling methodologies, and application of clinically validated thresholds. The model development phase achieved good performance metrics, with an AUC of 0.837, sensitivity of 0.86, and specificity of 0.79 during the clinical validation attempts (Fig. 1). The deployment architecture successfully materialized into a web-based clinical decision support tool, featuring real-time risk prediction capabilities and user-friendly interface elements through Streamlit implementation (https://readmission-prediction.streamlit.app/).

XAI

SHAP for global contribution across entire dataset

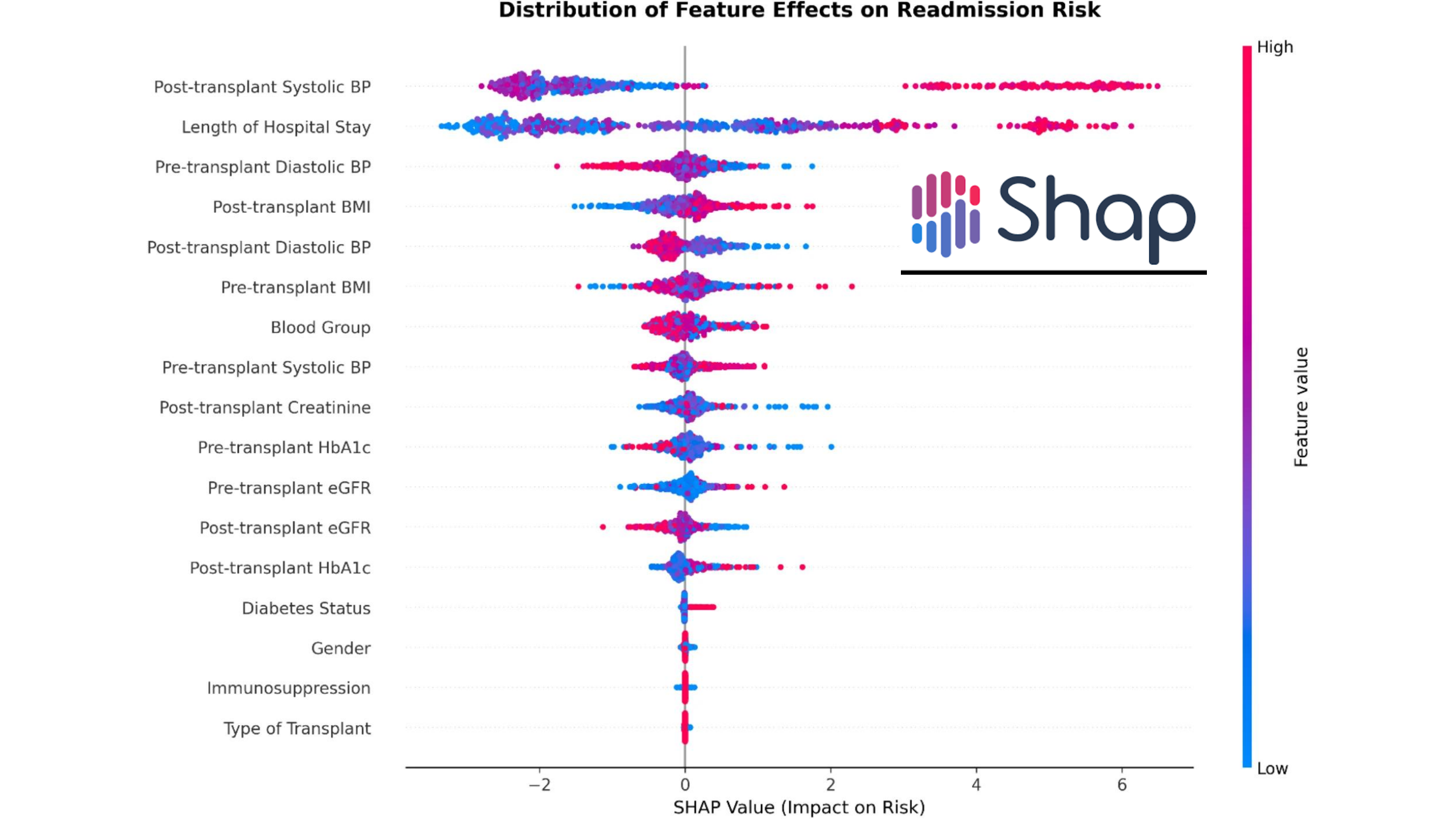
LIME for individual case prediction

"Why Should I Trust You?"
Explaining the Predictions of Any Classifier

Marco Tulio Ribeiro
University of Washington
Seattle, WA 98105, USA
marcotcr@cs.uw.edu

Sameer Singh
University of Washington
Seattle, WA 98105, USA
sameer@cs.uw.edu

Carlos Guestrin
University of Washington
Seattle, WA 98105, USA
guestrin@cs.uw.edu


Distribution of Feature Effects on Readmission Risk
Post-transplant Systolic BP
Length of Hospital Stay
Pre-transplant Diastolic BP
Post-transplant BMI
Post-transplant Diastolic BP
Pre-transplant BMI
Blood Group
Pre-transplant Systolic BP
Post-transplant Creatinine
Pre-transplant HbA1c
Pre-transplant eGFR
Post-transplant eGFR
Post-transplant HbA1c
Diabetes Status
Gender
Immunosuppression
Type of Transplant
Shap
−2
0
2
4
6
SHAP Value (Impact on Risk)
High
Feature value
Low

Clinical Predictors of Readmission Risk

Factor Importance Analysis Based on Machine Learning Model

Primary Risk Factors

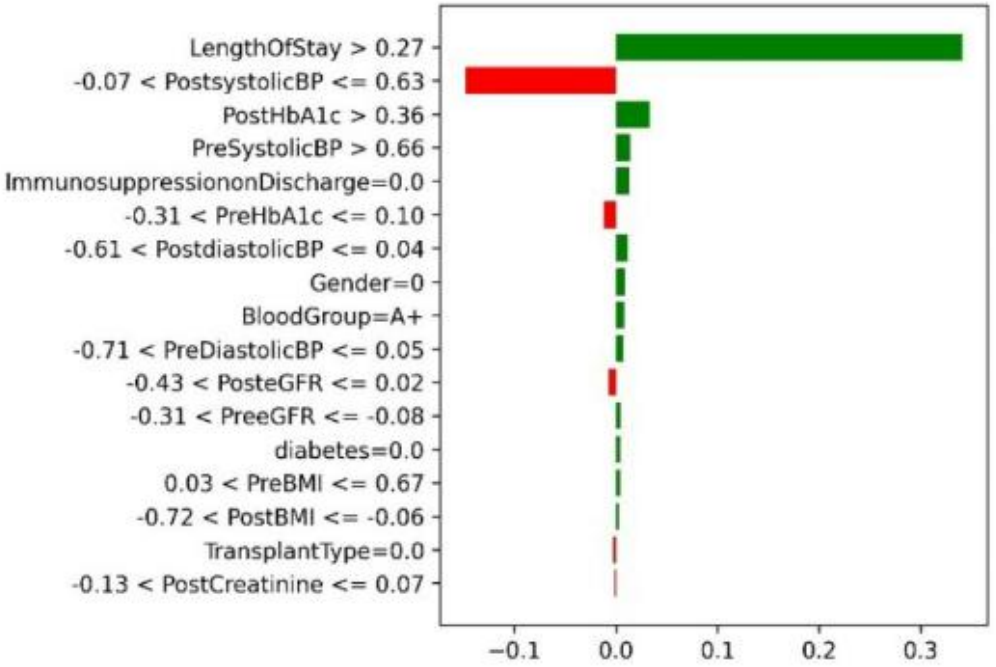
Secondary Risk Factors

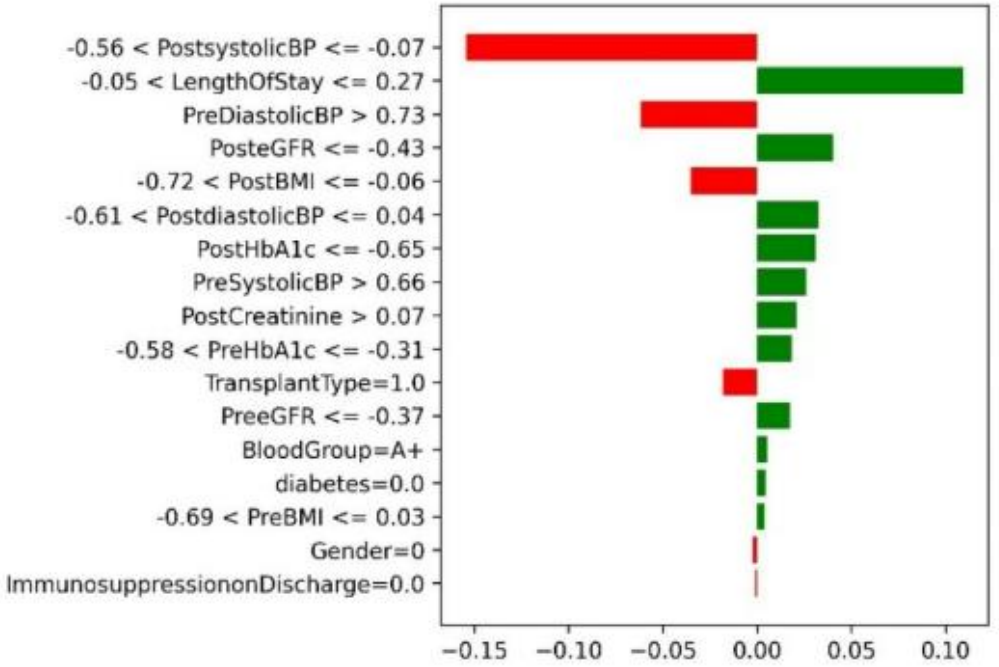
Additional Risk Factors

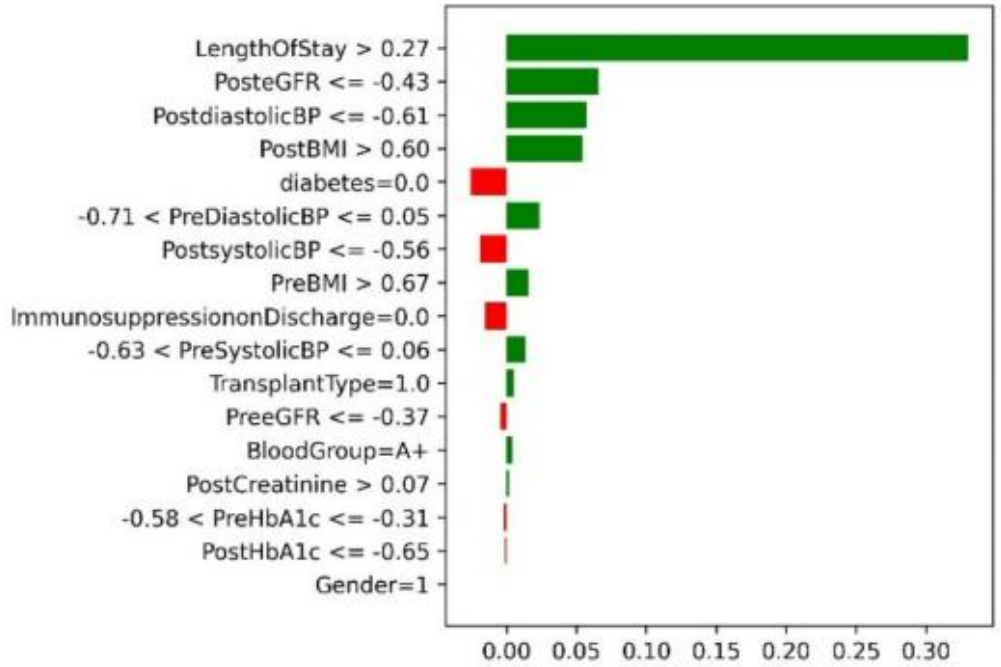
- Pre-transplant HbA1c (2.3%)
- Post-transplant eGFR (2.1%)
- Pre-transplant Systolic BP (2.1%)
- Pre-transplant eGFR (2.0%)
- Post-transplant HbA1c (1.7%)
- Gender (0.3%)
- Diabetes Status (0.1%)
- Other factors (0.1%)

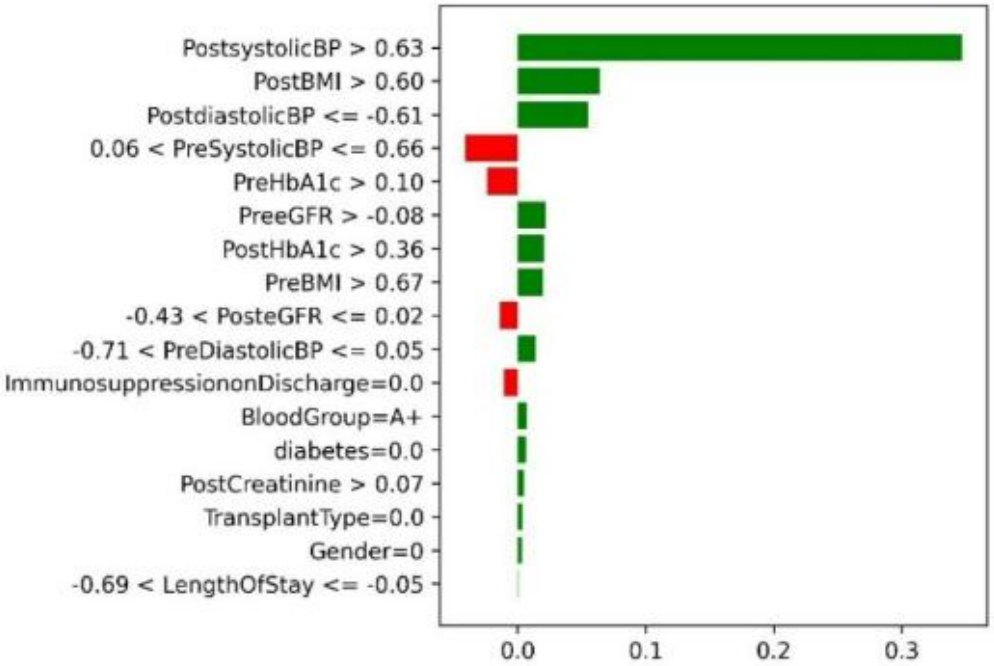
Key Finding: Length of stay and post-transplant blood pressure monitoring are critical for readmission risk assessment

Random case with LIME









Algorithm selection and training

Table 2 Model performance comparison between different algorithms

Metric	Random forest	XGBoost	Gradient boosting	Logistic regression	SVM	KNN
Accuracy	0.765 ± 0.034	0.789 ± 0.031	0.796 ± 0.050	0.740 ± 0.031	0.760 ± 0.041	0.707 ± 0.005
Precision	0.549 ± 0.115	0.590 ± 0.048	0.629 ± 0.090	0.183 ± 0.186	0.000 ± 0.000	0.293 ± 0.124
Recall	0.210 ± 0.048	0.447 ± 0.124	0.388 ± 0.129	0.027 ± 0.025	0.000 ± 0.000	0.119 ± 0.015
F1-Score	0.296 ± 0.038	0.494 ± 0.063	0.469 ± 0.105	0.047 ± 0.043	0.000 ± 0.000	0.164 ± 0.034
ROC-AUC	0.731 ± 0.029	0.799 ± 0.030	0.837 ± 0.035	0.604 ± 0.088	0.534 ± 0.054	0.550 ± 0.046

Note: Values are presented as Mean ± Standard Deviation across 5-folds

Algorithm selection and training

Table 2 Model performance comparison between different algorithms

Metric	Random forest	XGBoost	Gradient boosting	Logistic regression	SVM	KNN
Accuracy	0.765 ± 0.034	0.789 ± 0.031	0.796 ± 0.050	0.740 ± 0.031	0.760 ± 0.041	0.707 ± 0.005
Precision	0.549 ± 0.115	0.590 ± 0.048	0.629 ± 0.090	0.183 ± 0.186	0.000 ± 0.000	0.293 ± 0.124
Recall	0.210 ± 0.048	0.447 ± 0.124	0.388 ± 0.129	0.027 ± 0.025	0.000 ± 0.000	0.119 ± 0.015
F1-Score	0.296 ± 0.038	0.494 ± 0.063	0.469 ± 0.105	0.047 ± 0.043	0.000 ± 0.000	0.164 ± 0.034
ROC-AUC	0.731 ± 0.029	0.799 ± 0.030	0.837 ± 0.035	0.604 ± 0.088	0.534 ± 0.054	0.550 ± 0.046

Note: Values are presented as Mean ± Standard Deviation across 5-folds

No calibration performance report

What threshold used for discrimination performance?

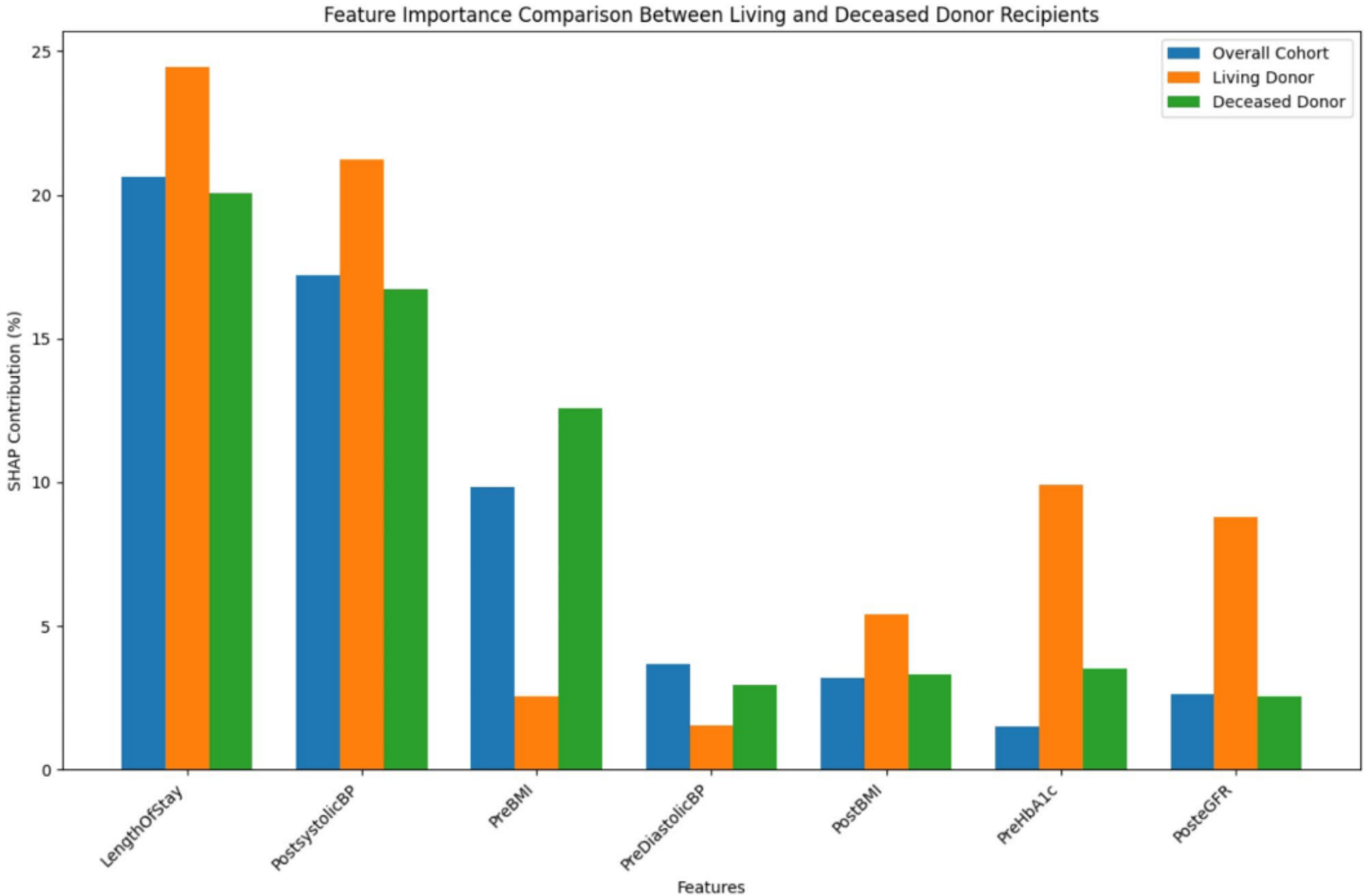
Where is PR-AUC?

Algorithm selection and training

Table 3 Subgroup analysis of model performance and feature importance in living vs. Deceased donor transplantation

Characteristic	Overall cohort (*n* = 588)	Living donor recipients (*n* = 500)	Deceased donor recipients (*n* = 87)	*P*-value*
Clinical outcomes:				
Readmission rate within 30 days (%)	88.9%	88.4%	92.0%	0.430
Mean hospital length of stay (days)	4.0 ± 4.1	3.9 ± 4.0	4.6 ± 4.6	0.259
Graft rejection episodes	110 (18.7%)	69 (13.8%)	41 (47.1%)	**0.000**
Model Performance Metrics:				
AUC (95% CI)	0.837 (0.802–0.872)	0.787 (0.738–0.836)	0.762 (0.685–0.839)	N/A
Sensitivity	0.388	0.402	0.412	N/A
Specificity	0.72	0.69	0.71	N/A
Accuracy	0.796 ± 0.050	0.778 ± 0.061	0.783 ± 0.058	N/A
Precision	0.629 ± 0.090	0.654 ± 0.082	0.643 ± 0.097	N/A
F1-Score	0.469 ± 0.105	0.453 ± 0.101	0.498 ± 0.112	N/A
Feature Importance (SHAP Contribution %):				
Length of hospital stay	20.6%	24.5%	20.1%	N/A
Post-transplant systolic BP	17.2%	21.2%	16.7%	N/A
Pre-transplant BMI	9.8%	2.6%	12.6%	N/A
Pre-transplant diastolic BP	3.7%	1.5%	2.9%	N/A
Post-transplant BMI	3.2%	5.4%	3.3%	N/A
Pre-transplant HbA1c	1.5%	9.9%	3.5%	N/A
Post-transplant eGFR	2.6%	8.8%	2.6%	N/A

Notes: * p-values compare living vs. deceased donor groups using appropriate statistical tests: t-test or Mann-Whitney U test for continuous variables depending on distribution normality; Chi-square or Fisher's exact test for categorical variables. N/A indicates comparison was not performed for this metric


Feature Importance Comparison Between Living and Deceased Donor Recipients
Overall Cohort
Living Donor
Deceased Donor
SHAP Contribution (%)
0
5
10
15
20
25
LengthOfStay
PostsystolicBP
PreBMI
PreDiastolicBP
PostBMI
PreHbA1c
PosteGFR
Features

Discussion

- From a clinical perspective, the model's well performing capabilities (AUC 0.837) **translate to practical utility in identifying high-risk patients** who may benefit from optimized and focused monitoring and early intervention.
- Our observed **readmission rate of 88.9% significantly exceeds previously reported ranges of 18–47%** in transplant literature
  - Our follow-up protocol involves **intensive post-transplant monitoring with a low threshold for readmission**, especially for the laboratory abnormalities that might be managed outpatient elsewhere.
  - **High proportion of living donor recipients** (85.2%) in our cohort may paradoxically **lead to more aggressive intervention for minor complications** given the elective nature of these transplants and heightened attention to outcomes.

Discussion

- Our dual-approach interpretability framework using **SHAP and LIME analyses transforms the complex machine learning outputs into actionable clinical key points** and insights for the readers
- For physicians, this means that the **model not only predicts readmission risk** but also **explains why specific patients are classified as high-risk, enabling more informed clinical decision-making.**
- While our **model identified length of hospital stay and post-transplant systolic blood pressure as primary statistical predictors**, these associations **should not be interpreted as directly modifiable intervention** targets without further investigation.

Discussion

- Our findings should therefore **guide risk stratification and resource allocation** rather than suggesting that **artificial manipulation of these parameters** (e.g., prematurely discharging patients or aggressively lowering blood pressure) **would necessarily reduce readmission risk**. Future interventional studies are required to determine which factors, if any, represent causal, modifiable targets for reducing readmission risk.

Transplant Readmission Risk Predictor

Patient Information

Gender

Male

Blood Group

A

Transplant Type

Living

Living Donor Model Performance:
AUC = 0.736 (95% CI: 0.576-0.897)
Sensitivity = 0.79, Specificity = 0.69

Diabetes Status

Yes

Pre-transplant Measurements

Pre-transplant BMI (kg/m²)

25.00

Post-transplant Parameters

Post-transplant BMI (kg/m²)

25.00

Post-transplant HbA1c (%)

5.70

Post-transplant Systolic BP (mmHg)

120

Post-transplant Diastolic BP (mmHg)

80

Post-transplant eGFR (mL/min)

60.00

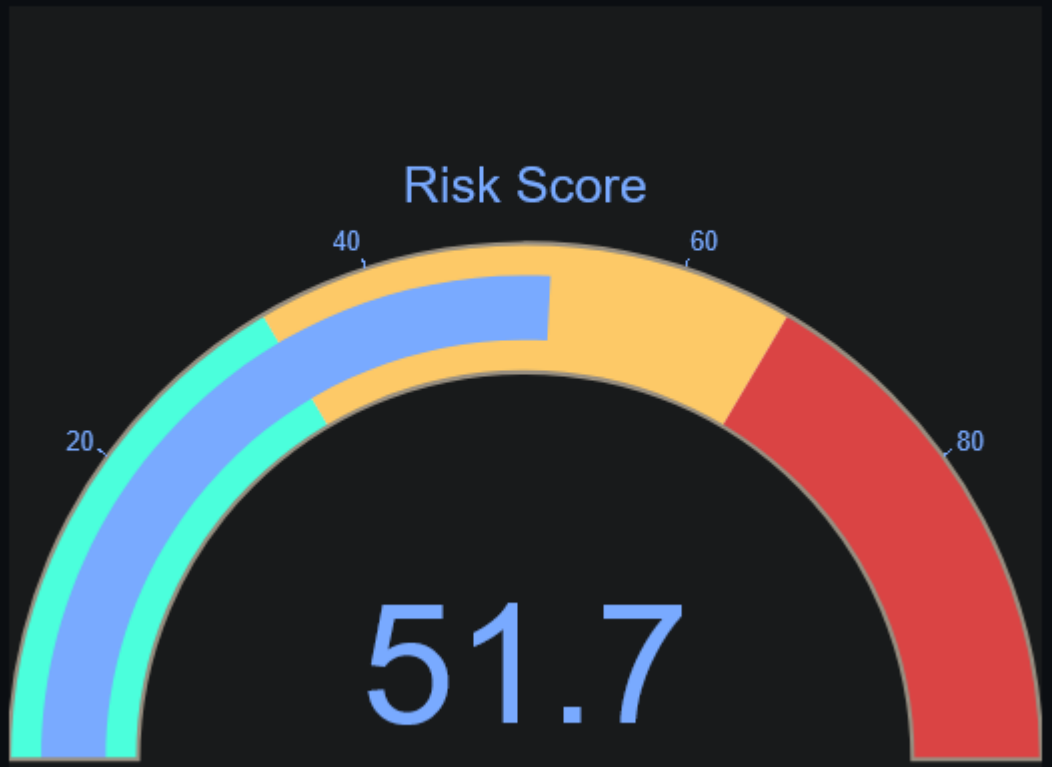
Post-transplant Creatinine (mg/dL)

1.20

Length of Stay (days)

5

Risk Assessment Results



Readmission Risk

🟡 **Medium Risk**

Probability: 51.7%

Donor Type: Living

✓ **Graft Rejection Risk: 13.8%**
Living donor recipients have lower rejection rates (13.8% vs 47.1%)

Recommended Actions

- 📋 Regular monitoring
- 🏥 Follow-up within 1 week
- 📞 Twice-weekly check-ins
- ⚕️ Review medications

Clinical Interpretation

Key Risk Factors:

- Length of Stay: 5 days
- Post-transplant Systolic BP: 120 mmHg
- Pre-transplant HbA1c: 5.7%
- Post-transplant eGFR: 60.0 mL/min

Donor-Specific Context:

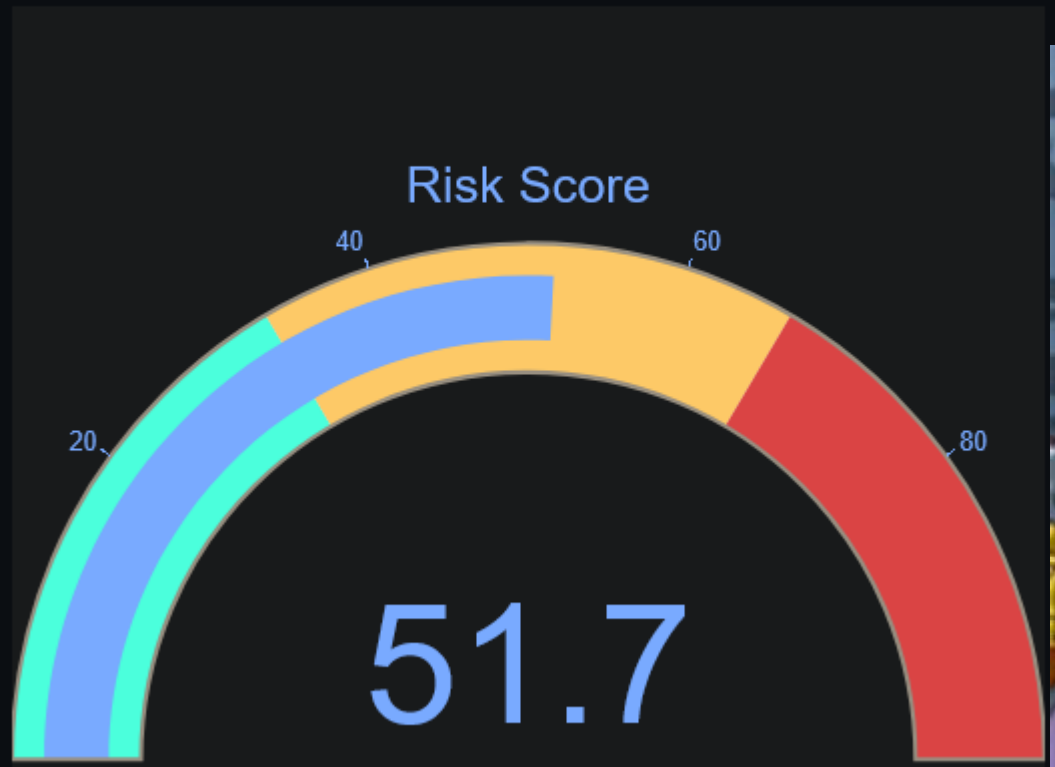
- Living donor transplant (13.8% rejection risk)
- ⚠️ HbA1c and eGFR are stronger predictors for living donors
- ✓ Normal stay
- ✓ Controlled BP

Model Confidence:

- Prediction confidence: 51.7%
- Based on 5 days of monitoring
- Living donor model: AUC 0.736
-

Present probability but no calibration performance report

Risk Assessment Results



Clinical Interpretation

Key Risk Factors:

- Length of Stay: 5 days
- Post-transplant Systolic BP: 120 mmHg
- Pre-transplant HbA1c: 5.7%
- Post-transplant eGFR: 60.0 mL/min

Donor-Specific Cor

- Living donor transplant (13.8% rejection risk)
- ⚠️ HbA1c and eGFR are stronger predictors for living donors
- ✓ Normal stay
- ✓ Controlled BP

- Prediction confidence: 51.7%
- Based on 5 days of monitoring
- Living donor model: AUC 0.736
-

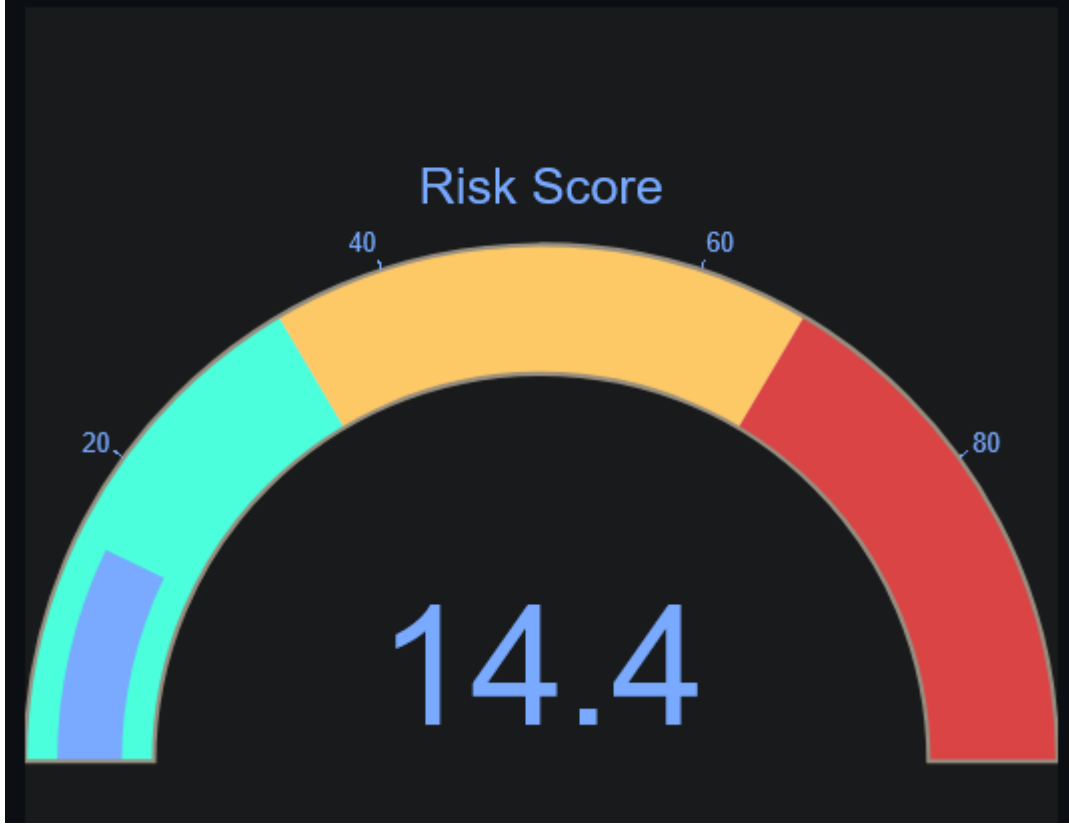
This tool is intended to assist clinical decision-making and should not replace professional medical judgment.

Model performance varies by donor type (Living: AUC 0.736, Deceased: AUC 0.708).

For emergency situations, please contact your healthcare provider immediately.


Length of Stay (days)
5
−
+
Length of Stay (days)
6
−
+

Risk Assessment Results



Readmission Risk

🟢 **Low Risk**

Probability: 14.4%

Donor Type: Living

✓ **Graft Rejection Risk: 13.8%**
Living donor recipients have lower rejection rates (13.8% vs 47.1%)

Recommended Actions

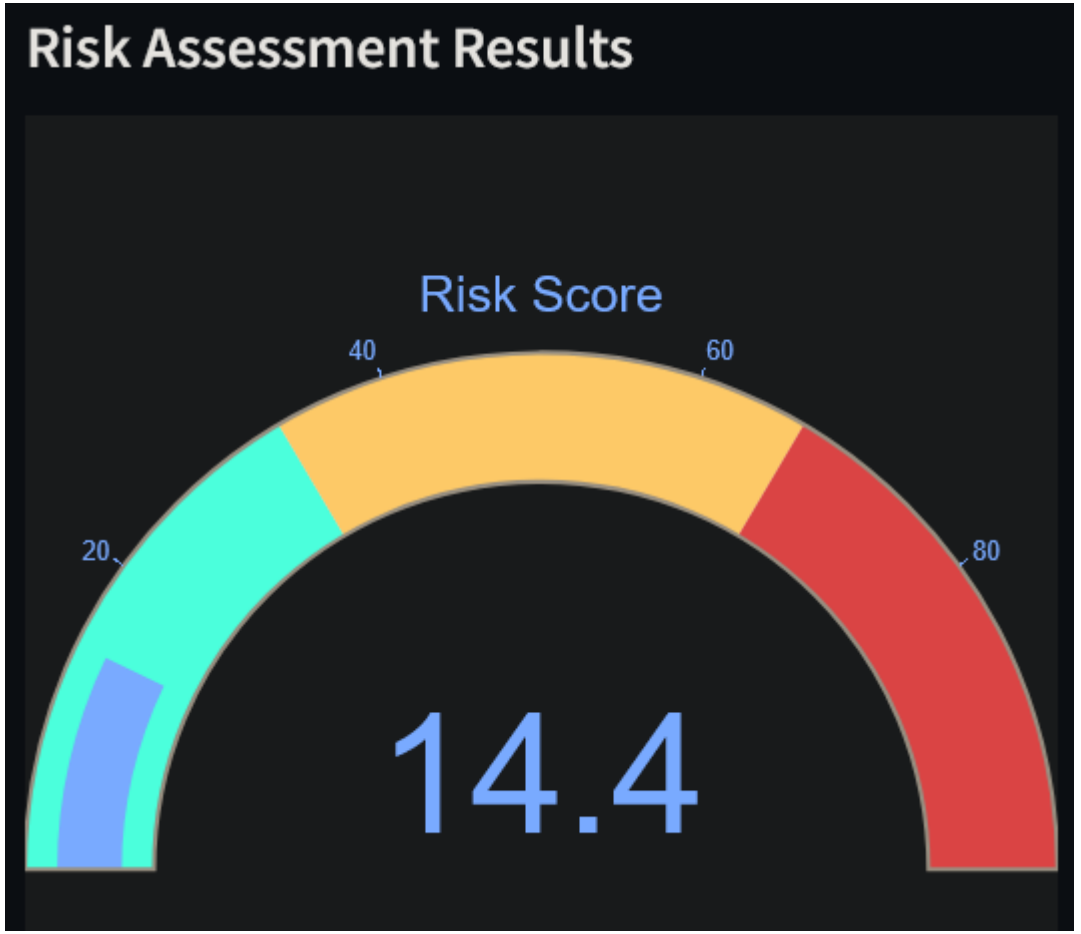
- 📋 Routine monitoring
- 📅 Standard schedule
- 📞 Weekly check-ins
- ✅ Standard care

Length of Stay (days)

5

Length of Stay (days)

6



Length of Stay (days)

7

Risk Assessment Results

My Critisms

- **Primary endpoint is meaningless.**
  - Why 30 days? Why include all hospital contact? Why not use time-to-event model?
- **Data leakage during feature selection.**
- **No calibration performance, no clinical or utility justification for threshold used.**
  - The dataset used already has low threshold for admission due to institutional protocol.

    Admission does not means “bad outcome”, but rather the admission is to “prevent bad outcome”, so admission is a meaningless label.
    - Exceedingly high admission prevalence compared to other place necessitate calibration.
- **Only association, not causation -> cannot be used to modify protocol**
- **No sample size justification.**
- **Is 5-fold CV enough to correct for optimism?**
  - Most simulation suggest **repeated k-fold** at least. Or **bootstrap**.

Recommended Actions

-  Regular monitoring
-  Follow-up within 1 week
-  Twice-weekly check-ins
-  Review medications

This recommendation cannot be made

The significant improvement in predictive capabilities offered by our model, combined with its clinical interpretability, positions it as a valuable addition to the transplant physician's toolkit. By providing quantifiable, evidence-based risk assessment, the n s clinical judgment in optimizing post-transplant care strategies. As transplant medicine continues to evolve, such AI-driven tools will become increasingly important in achieving improved patient outcomes through personalized care approaches.

It is not.

By providing quantifiable, evidence-based risk assessment, the model supports clinical judgment in optimizing post-transplant care strategies. evolve, such AI-driven tools will become increasingly im ent outcomes through personalized care approaches.

This model probably will make the decision worse.

From previous journal club

Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission

Rich Caruana
Microsoft Research
rcaruana@microsoft.com

Yin Lou
LinkedIn Corporation
ylou@linkedin.com

Johannes Gehrke
Microsoft
johannes@microsoft.com

Paul Koch
Microsoft Research
paulkoch@microsoft.com

Marc Sturm
NewYork-Presbyterian Hospital
mas9161@nyp.org

Noémie Elhadad
Columbia University
noemie.elhadad@columbia.edu

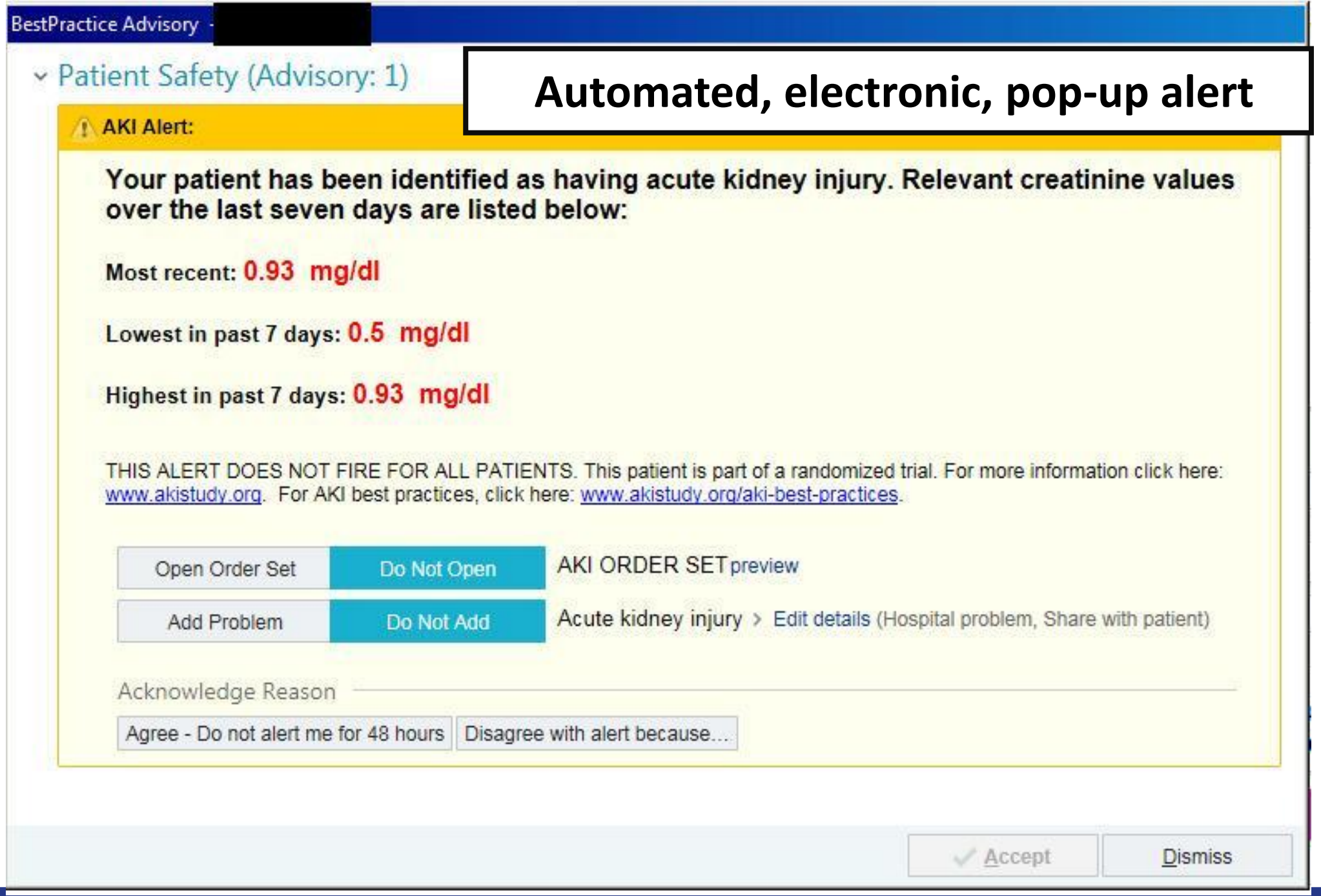
CEHC study faced the same problem in the mid 90s

https://dl.acm.org/doi/10.1145/2783258.2788613

Electronic health record alerts for acute kidney injury: multicenter, randomized clinical trial

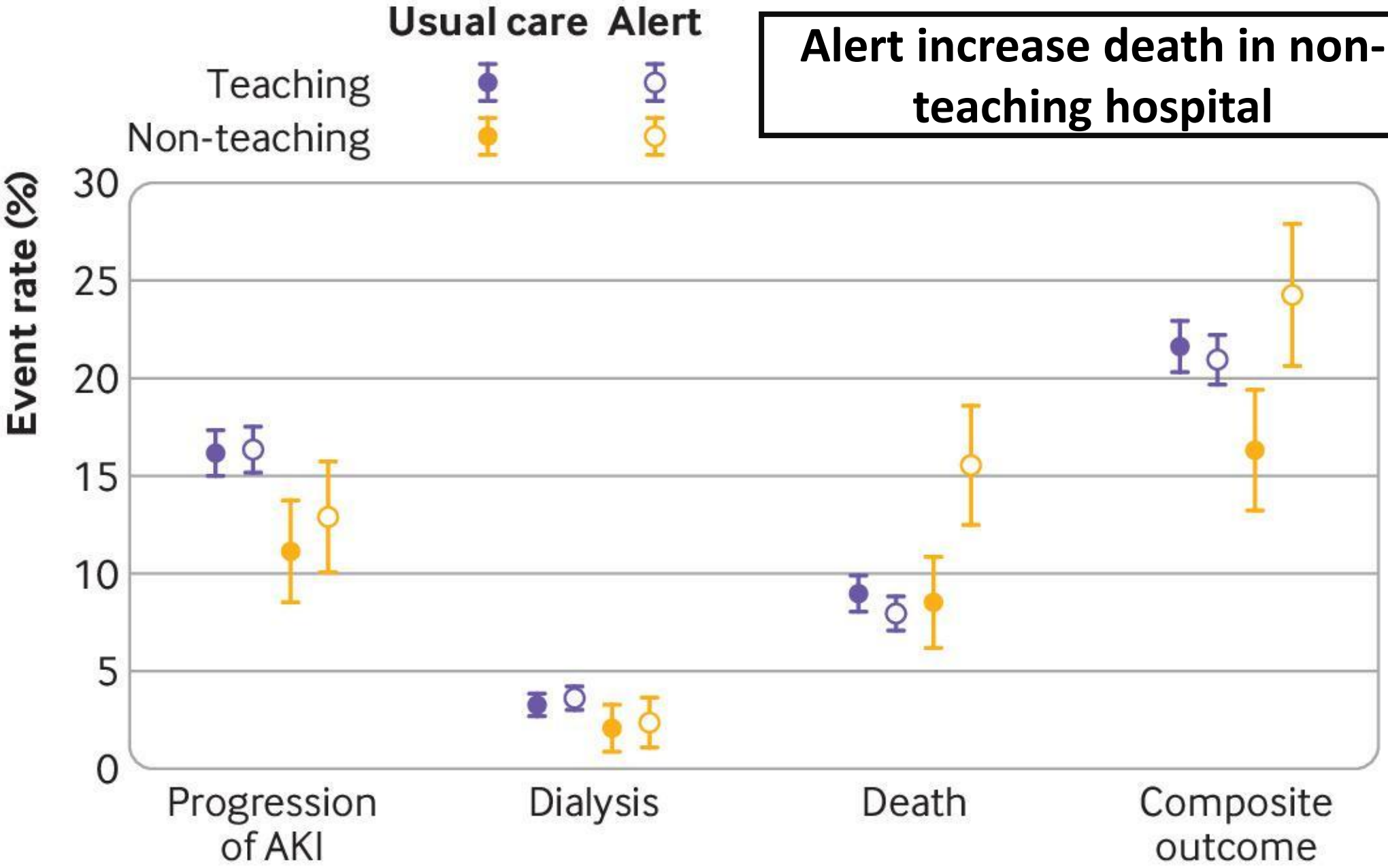
F Perry Wilson,[1,2] Melissa Martin,[1,2] Yu Yamamoto,[1,2] Caitlin Partridge,[3] Erica Moreira,[3] Tanima Arora,[1,2] Aditya Biswas,[1,2] Harold Feldman,[4] Amit X Garg,[5] Jason H Greenberg,[2,6] Monique Hinchcliff,[7] Stephen Latham,[8] Fan Li,[9] Haiqun Lin,[10] Sherry G Mansour,[1,2] Dennis G Moledina,[1,2] Paul M Palevsky,[11] Chirag R Parikh,[12] Michael Simonov,[2] Jeffrey Testani,[13] Ugochukwu Ugwuowo[1,2]

Nat Sirirutbunkajorn
Radiation oncologist, Ramathibodi hospital
Department of Clinical Epidemiology and Biostatistics, Faculty of Medicine Ramathibodi Hospital, Mahidol university
Nut19012537@gmail.com



https://www.bmj.com/content/372/bmj.m4786

Alert increase death in non-teaching hospital



https://www.bmj.com/content/372/bmj.m4786

It was upon inspection of the final results that we got worried. Two of our six medical centers, the two non-teaching hospitals, had results that were worse in the alert than the usual care arm. Specifically, the mortality rate was higher in the alert arm.

Needless to say, we immediately began a deep dive into the data—first confirming that somehow we had not flipped our randomization variable and to ensure balance between the randomization groups—and then going deeper to evaluate for potential mediators of the effect. Could patients in the alert arm have received inappropriate fluid resuscitation? No sign of that. Did they have a lower rate of key diagnostic tests (in a misguided attempt to avoid contrast perhaps)? No signal. In the end, frustratingly, we could attribute the harm to no specific process.

We informed the IRB and the study hospitals, who launched their own investigations. We manually adjudicated every study death in those hospitals looking for a common theme and found nothing. Patients died for reasons many patients die—heart failure, malignancy, sepsis. Aside from the greater *number* of deaths in the alert arm, there was no sign that any mechanism of death was distributed unevenly between the groups.

https://blogs.bmj.com/bmj/2021/01/18/the-challenge-of-minimal-risk-in-e-alert-trials/

We were left with an unsettling situation. The outcome of this work demonstrated the need for this type of research. Alerts are increasingly marketed to hospitals and adopted with the intent to improve care. But do they improve care? Indeed, are they at the very least benign? We accept that the impact of alerts needs to be studied, but how? We had described this study as minimal risk because, we thought, it must be. But our data suggest that assumption may not have been accurate in all cases. To be sure, the effects we saw may be due to the vicissitudes of chance—the analysis of non-teaching hospitals was a sub-group analysis after all. But what if we are seeing an example of heterogeneity of treatment effect? A phenomenon whereby the impact of an intervention differs among different groups due to a variety of factors upstream and downstream of the intervention—many of which may not be easily measured.

https://blogs.bmj.com/bmj/2021/01/18/the-challenge-of-minimal-risk-in-e-alert-trials/