# Tips for clinically meaningful machine learning in healthcare

Sunny Lou, MD PhD

2025.09.09

# Great interest in ML for health

THE WALL STREET JOURNAL.

SUBSCRIBE | SIGN IN

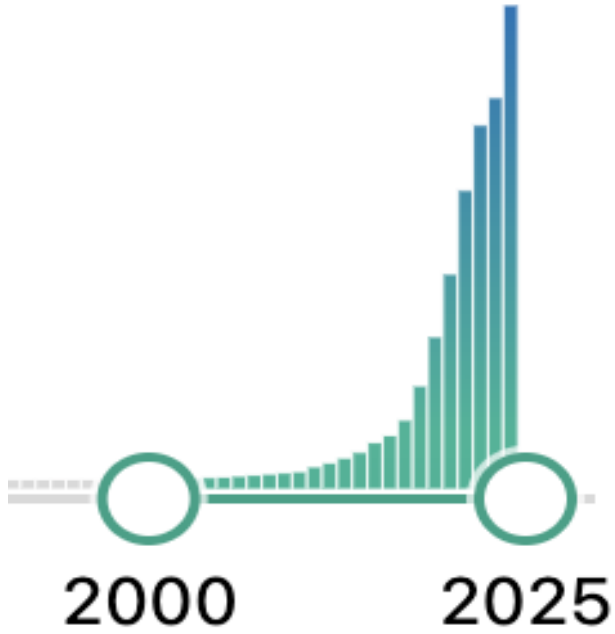
TECH | KEYWORDS

## The AI Doctor Will See You Now

Artificial intelligence comes to the doctor's office, helping identify disease, monitor heart activity, stave off seizures



**200k+ results!!**



# Relatively few success stories

**Original Investigation** | Health Informatics

September 29, 2022

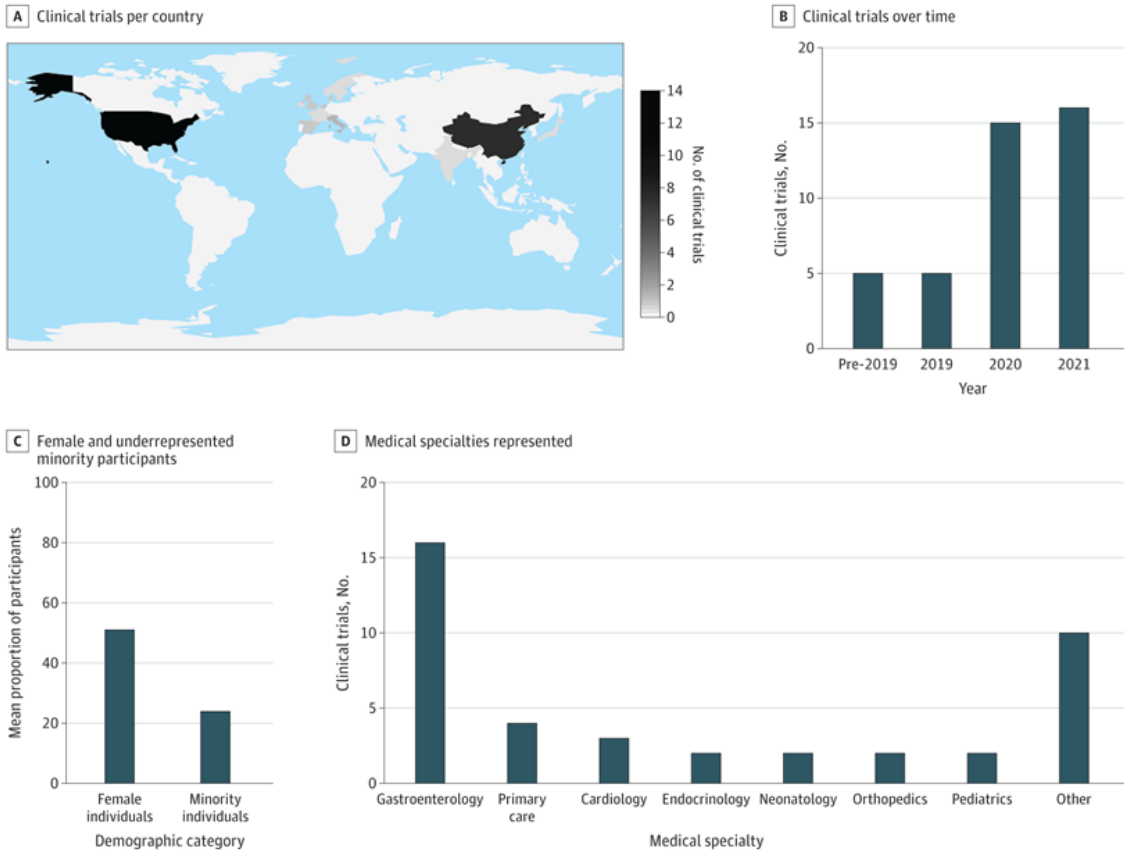
## Randomized Clinical Trials of Machine Learning Interventions in Health Care
## A Systematic Review

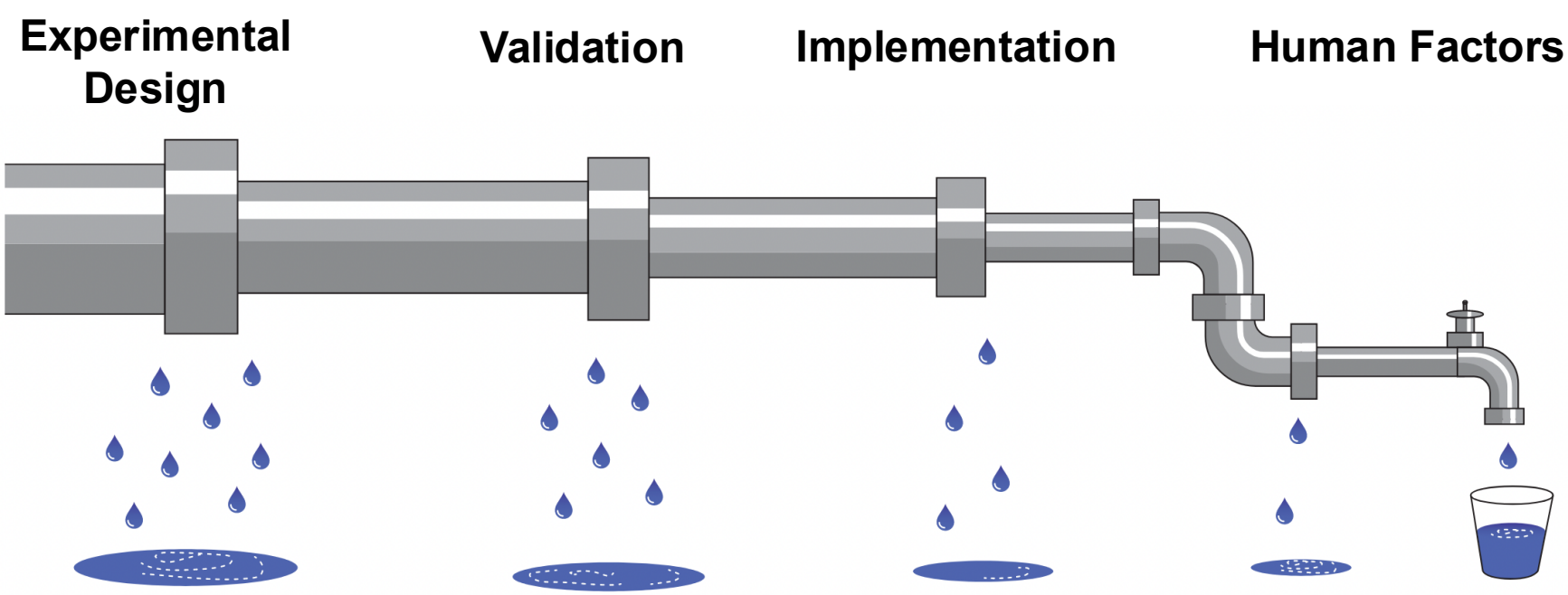
Deborah Plana, BS[1]; Dennis L. Shung, MD, PhD[2]; Alyssa A. Grimshaw, MSLIS[3]; et al

» Author Affiliations | Article Information

*JAMA Netw Open.* 2022;5(9):e2233946. doi:10.1001/jamanetworkopen.2022.33946

**41 trials through 10/15/21**






| Not fit for purpose | No validation | No implementation | Not adopted |
|---|---|---|---|
| Developed on wrong patient population | Lack of data or incentive to pursue validation studies | No impact on decision making or patient (health) outcomes | Prediction (perceived as) not useful |
| Expensive or non-available predictors | Incompletely reported prediction model | No software developed to implement and use the model | Predictions not trusted |
| Time intensive to use model | Poorly developed or overfitted model | Requirements for adherence to (medical device) regulations | Model not transparent enough, or no tools available to enhance its use in practice |
| Outcome measured unreliably | Proprietary model code | Cost(-effectiveness) of using proprietary model | Model (perceived as) outdated |

**FIGURE 1** Leaky prognostic model adoption pipeline. Examples of reasons for failed prediction model adoption in clinical practice.

Royen et al (2022) Eur Resp J

# Tip 1: Make the prediction actionable

Is the outcome you're predicting attached to a specific clinical decision?

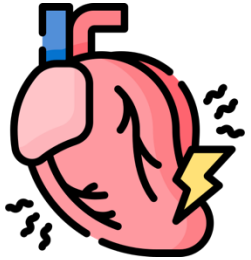
- Predict a clinician action
- Predict an outcome with obvious interventions

How do you expect clinicians to use the prediction?

- Who?
- At what time?







# Tip 2: Think about potential biases in your outcome

- How reliable is the label assignment?
- Could the label reflect systematic biases?

CORRESPONDENCE

Racial Bias in Pulse Oximetry Measurement

December 17, 2020
N Engl J Med 2020; 383:2477-2478
DOI: 10.1056/NEJMc2029240




**Dissecting racial bias in an algorithm used to manage the health of populations**

ZIAD OBERMEYER, BRIAN POWERS, CHRISTINE VOGELI, AND SENDHIL MULLAINATHAN Authors Info & Affiliations

*SCIENCE* • 25 Oct 2019 • Vol 366, Issue 6464 • pp. 447-453 • DOI: 10.1126/science.aax2342

# Tip 3: Choose your training dataset carefully

- Are all the patients you're hoping to apply your model to represented in your data?
- Do you care about generalizability?
- How is data quality?

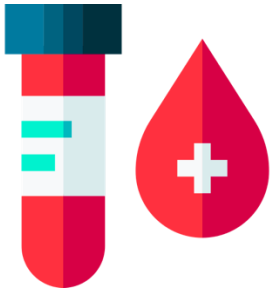
VA | U.S. Department of Veterans Affairs

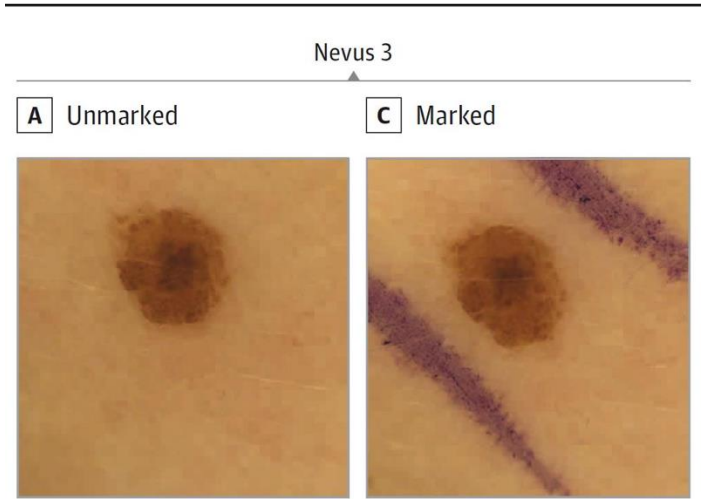


# Tip 4: Choose practical predictors

- Are the predictors available at the time you want your model to run?
- Are the predictors easily available?
- Label leakage

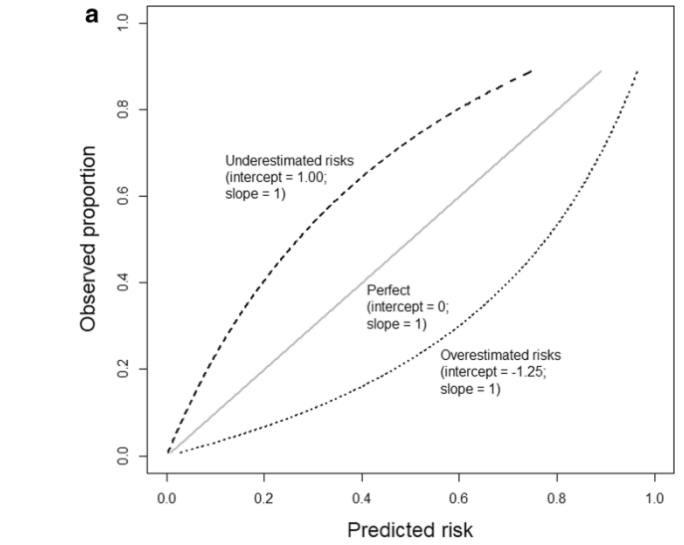
**ICD-10**

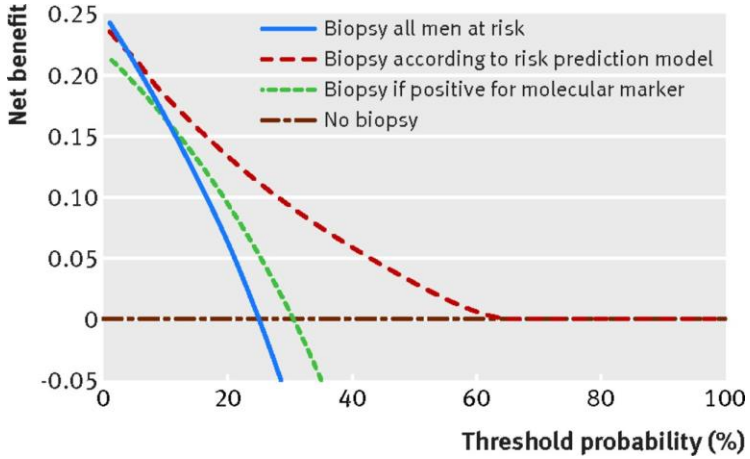




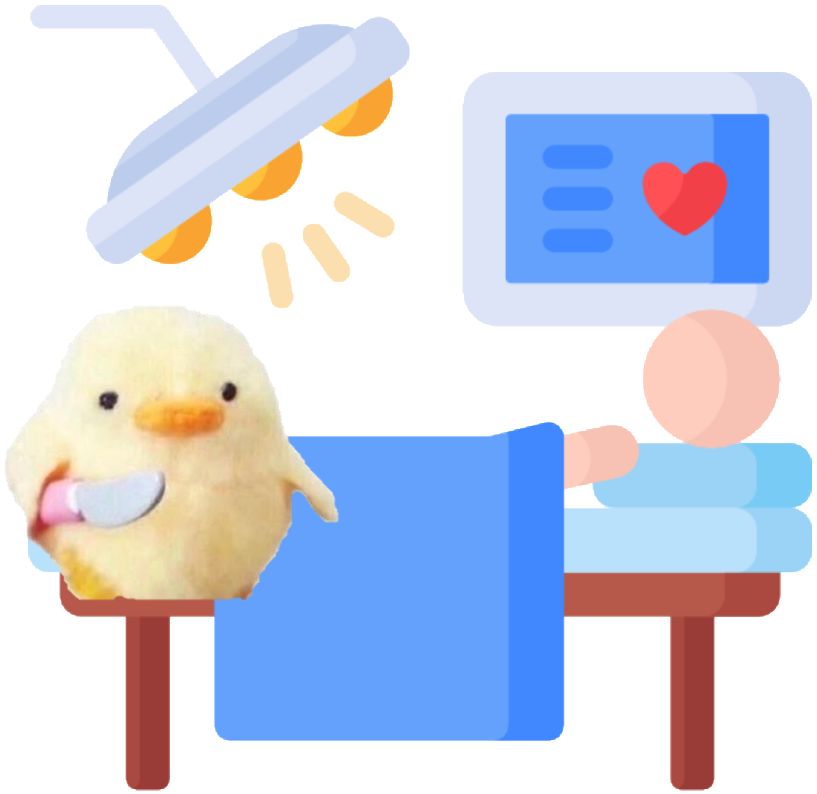
# Tip 5: Measure clinically meaningful performance

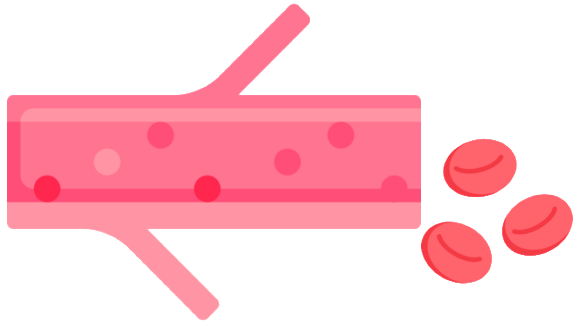
- AUROC / AUPRC are threshold-free
- If used for a specific decision: choose a threshold according to clinical needs
- If presenting risks to clinicians: Measure calibration
- Consider net benefit or comparison to existing clinical calculators

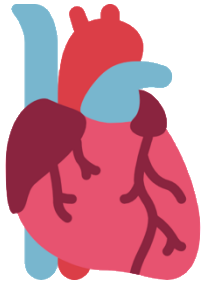


# Blood transfusion can be life-saving in surgery





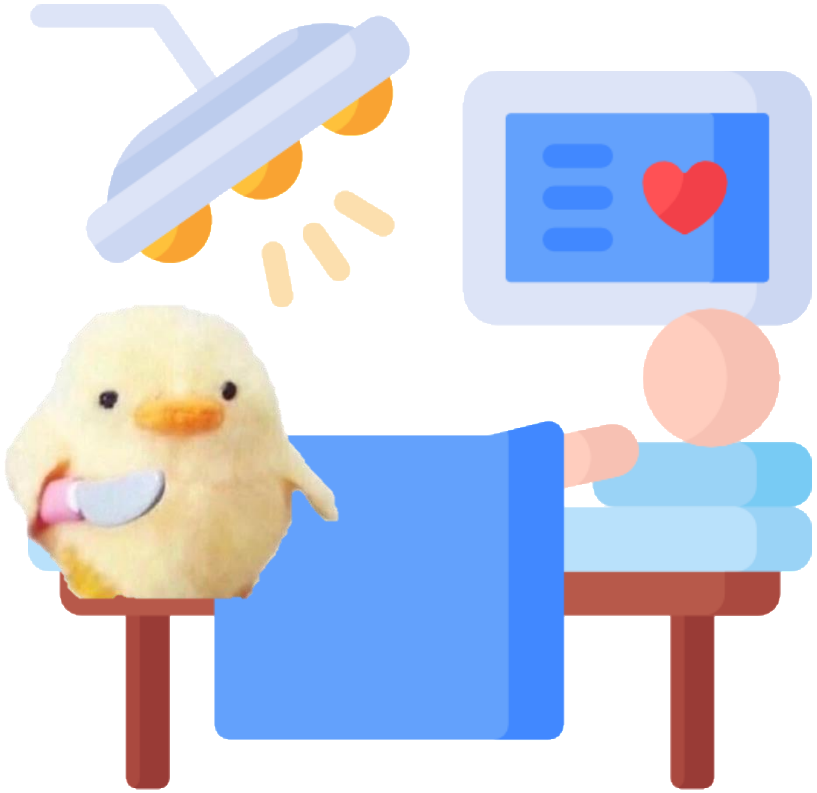


Provides volume to support cardiac output, perfusion pressure




Delivers oxygen to support tissue function

# Safe transfusion requires multiple steps





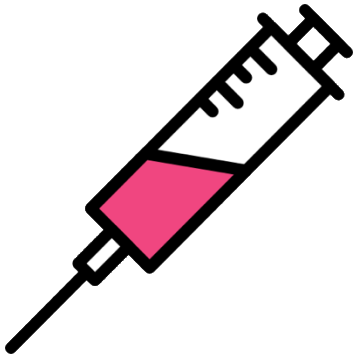
1. Identify correct blood type (~1h)
2. Find compatible unit (min - days)
3. Deliver unit to OR (15-30 min)

## Preoperative preparation

**Type and Screen** (T/S)

**Crossmatch**

# Excess preparation can be harmful



**Discomfort**

**Needle-stick injury**

**Patient cost**

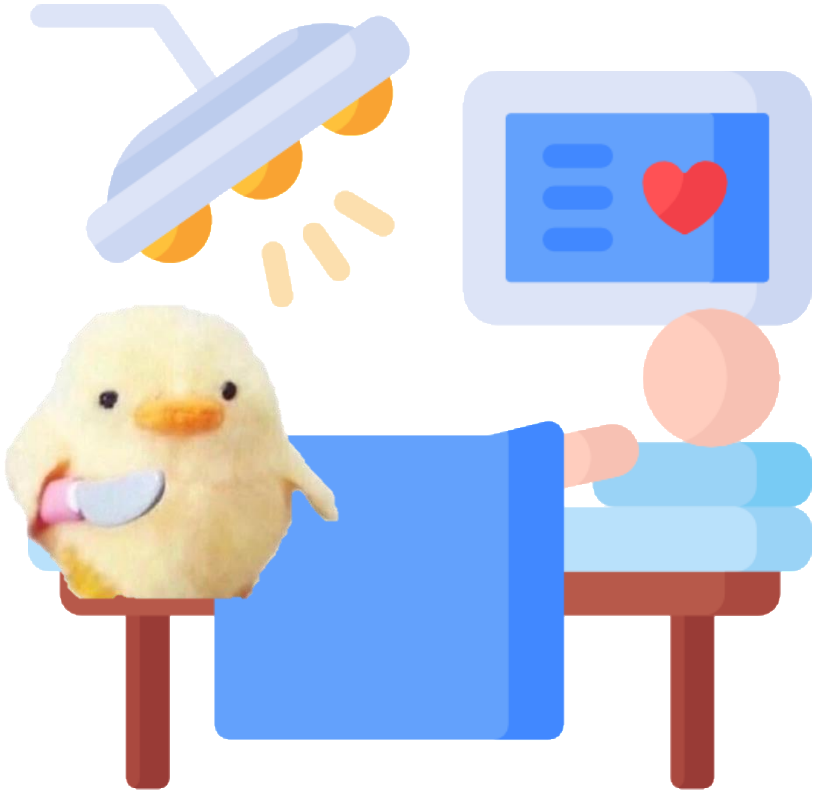


**Blood sustainability**

**Wasted labor**

**Healthcare system cost**

# Preoperative preparation: how to decide?





1. **Type and screen (T/S)** (~1h)
2. **Crossmatch** (min - days)
3. **Deliver to OR** (15-30 min)

How much blood loss will there be?

How much blood can the patient afford to lose?

How quickly will bleeding occur?

# Maximum surgical blood ordering schedule (MSBOS) determines T/S, crossmatch orders

**Procedure-specific**

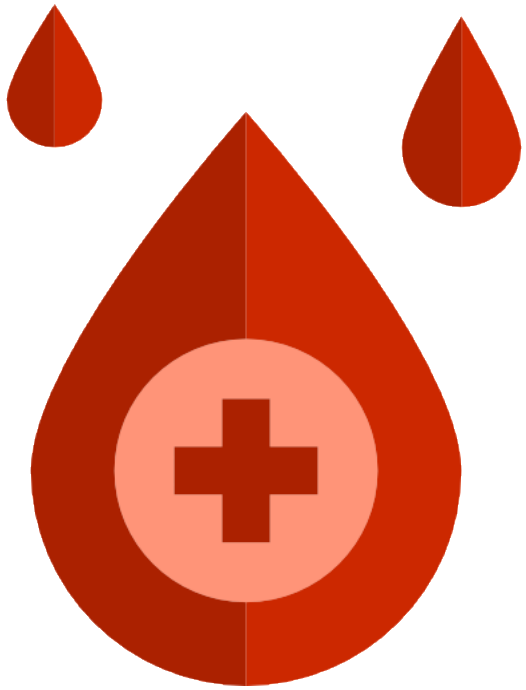
< 5% frequency of transfusion = No T/S

**Hospital-specific**

**NOT patient-specific**

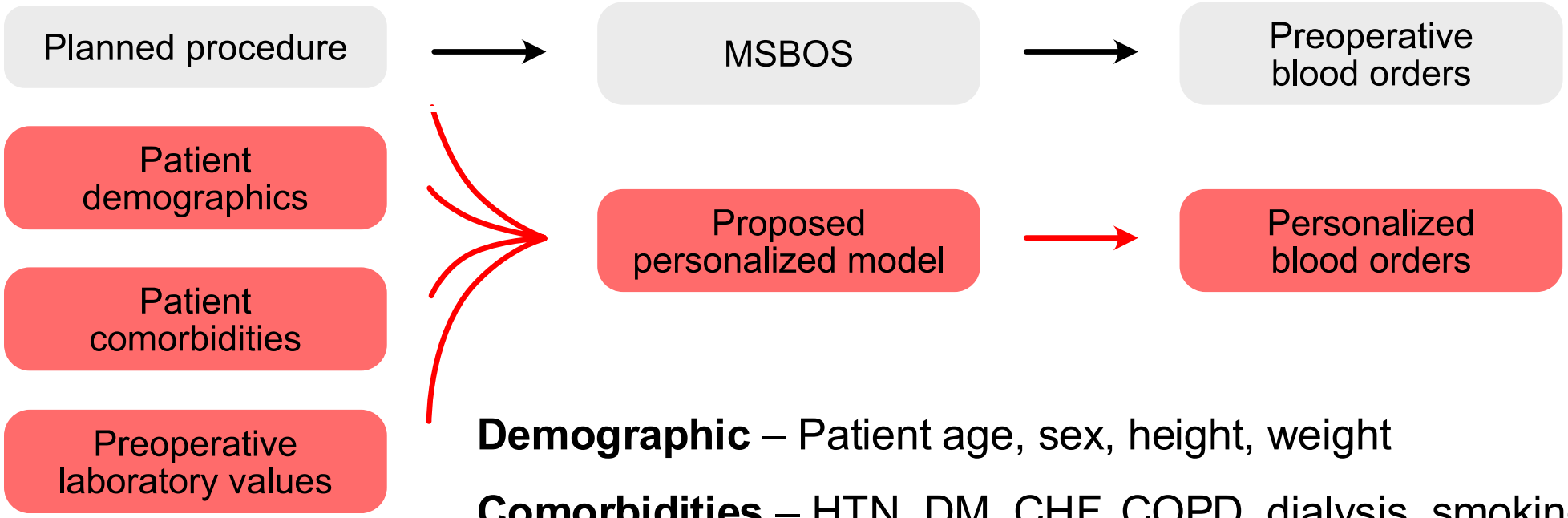
| *Service* | *Procedure* | *Routine blood bank order* |
|---|---|---|
| Cardiac Surgery | Pacemaker/AICD insertion (without extraction) | Type and Screen |
| | Pacemaker/AICD removal | Type and Cross (2) |
| | Coronary Bypass and Open Valve Cases | Type and Cross (4) |
| | Thoracic Aortic Surgery | Type and Cross (4) |
| | Transcatheter Aortic Valve Surgery (TAVR) | Type and Cross (4) |
| | ECMO | Type and Cross (6) |
| | Lung transplant surgery | Type and Cross (6) |
| | Redo Sternotomy for Heart Surgery | Type and Cross (6) |
| | Ventricular Assist Device (VAD) | Type and Cross (6) |
| | Heart transplant surgery | Type and Cross (10) |
| Cardiology | Transcatheter Mitral Valve Repair | Type and Screen |
| Electrophysiology | Pacemaker/AICD Insertion or Revision | None (Emergency Release Blood) |
| | Cardiac Ablation | None (Emergency Release Blood) |
| | Watchman device/Lariat Procedure (left atrial appendage exclusion | Type and Cross (2) |
| ENT | Facial Fractures | None (Emergency Release Blood) |
| | Laryngoscopy with Biopsy | None (Emergency Release Blood) |
| | Thyroidectomy/Parathyroidectomy | None (Emergency Release Blood) |
| | Tonsillectomy and/or Adenoidectomy | None (Emergency Release Blood) |
| | Hemiglossectomy | Type and Screen |

# Our goal



Build a **personalized** clinical decision support tool that can guide **pre-surgical blood orders**

# Approach for personalized risk




**Demographic** – Patient age, sex, height, weight

**Comorbidities** – HTN, DM, CHF, COPD, dialysis, smoking

**Preop labs** – Hg, Plt, INR, PTT, Na, Cr, albumin, bili

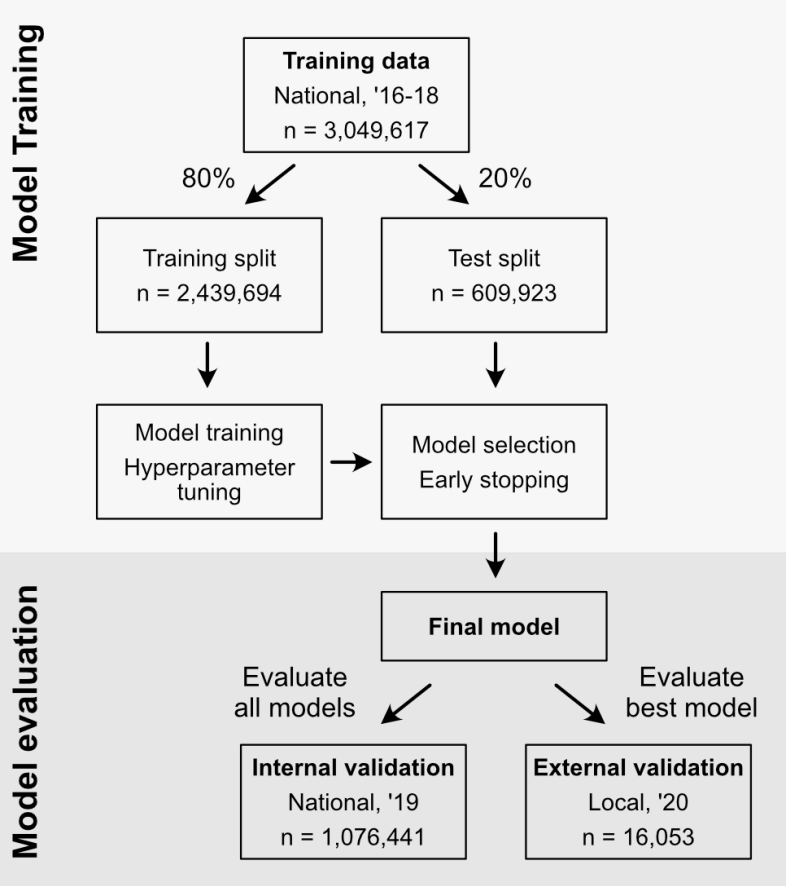
**Procedure** – procedure-specific risk, elective surgery

# Data Source: NSQIP

National Surgical Quality Improvement Program
American College of Surgeons

- Data submitted by ~700 hospitals across the United States, including community and academic hospitals
- Includes ~1 million surgical cases per year
- Data is entered by professional coders according to strict criteria
- Frequency of surgeries in NSQIP is not necessarily representative of overall frequency

# Experimental design




Choose variables easily extracted from EHR

Train models on NSQIP 2016-2018

Internal validation on NSQIP 2019

External validation on BJH 2020

# Cohort demographics

| Variable | Data Type | Training: (n = 3,049,617) | Internal Validation: (n = 1,076,441) | External Validation: (n = 16,053) |
|---|---|---|---|---|
| Transfused – Yes, n (%) | Binary | 73,313 (2.4%) | 23,205 (2.2%)* | 1,104 (6.7%)* |
| Age, mean (std) | Continuous | 57 (17) | 57 (17)* | 57 (16) |
| Height (in), mean (std) | Continuous | 66 (4) | 66 (4)* | 67 (4)* |
| Weight (lb), mean (std) | Continuous | 188 (51) | 186 (49)* | 193 (56)* |
| Gender – Male, n (%) | Binary | 1,316,142 (43%) | 454,517 (42%)* | 7,592 (47%)* |
| ASA Status – I, n (%) | Ordinal | 253,948 (8%) | 85,144 (7.9%)* | 658 (4%)* |
| II, n (%) | | 1,356,498 (45%) | 486,469 (45%)* | 5,962 (37%)* |
| III, n (%) | | 1,247,668 (41%) | 442,515 (41%)* | 7,504 (47%)* |
| IV, n (%) | | 178,613 (6%) | 57,950 (5%)* | 1,842 (11%)* |
| V, n (%) | | 5,413 (0.2%) | 1,905 (0.2%)* | 69 (0.4%)* |
| HTN – Yes, n (%) | Binary | 1,354,643 (44%) | 466,520 (43%)* | 7,227 (45%) |
| CHF – Yes, n (%) | Binary | 25,930 (1%) | 9,051 (1%) | 1,192 (7%)* |
| Smoking – Yes, n (%) | Binary | 516,646 (17%) | 165,551 (15%)* | 1,760 (11%)* |
| COPD – Yes, n (%) | Binary | 129,528 (4%) | 43,290 (4%)* | 1,337 (8%)* |
| Dialysis – Yes, n (%) | Binary | 39,343 (1%) | 11,459 (1%)* | 524 (3%)* |
| Diabetes – No, n (%) | Ordinal | 2,574,255 (84%) | 913,163 (85%)* | 13,192 (82%)* |
| Meds, n (%) | | 300,959 (10%) | 106,689 (10%)* | 1,881 (12%)* |
| Insulin, n (%) | | 174,402 (6%) | 56,589 (5%)* | 979 (6%)* |

| Variable | Data Type | Training: (n = 3,049,617) | Internal Validation: (n = 1,076,441) | External Validation: (n = 16,053) |
|---|---|---|---|---|
| Hematocrit, g/dL, mean (std) | Continuous | 39.6 (5.3) | 40.0 (5.3)* | 36.8 (6.8)* |
| Platelet, 1000/uL, mean (std) | Continuous | 251 (82) | 255 (82)* | 257 (110)* |
| INR, mean (std) | Continuous | 1.1 (0.3) | 1.1 (0.3)* | 1.2 (0.4)* |
| PTT, s, mean (std) | Continuous | 30 (8) | 31 (8)* | 34 (15)* |
| Creatinine, mg/dL, mean (std) | Continuous | 1.00 (0.86) | 0.98 (0.78)* | 1.35 (1.64)* |
| Sodium, mEq/L, mean (std) | Continuous | 139 (3) | 139 (3)* | 139 (4)* |
| Albumin, g/dL, mean (std) | Continuous | 4.0 (0.6) | 4.0 (0.6)* | 3.8 (0.8)* |
| Bilirubin, mg/dL, mean (std) | Continuous | 0.6 (0.7) | 0.6 (0.7)* | 0.7 (1.7)* |
| Elective surgery – Yes | Binary | 2,429,994 (80%) | 847,691 (79%)* | 10,686 (67%)* |
| Procedure-specific transfusion rate (%), mean (std) | Continuous | 2.4 (5.7) | 2.3 (5.5)* | 7.5 (15)* |

* indicates Bonferroni corrected $p < 0.05$ compared to training data

# Evaluation metrics for a surgical transfusion risk model designed to guide T&S order

**AUROC** – probability of rank-ordering 2 patients correctly in likelihood of transfusion

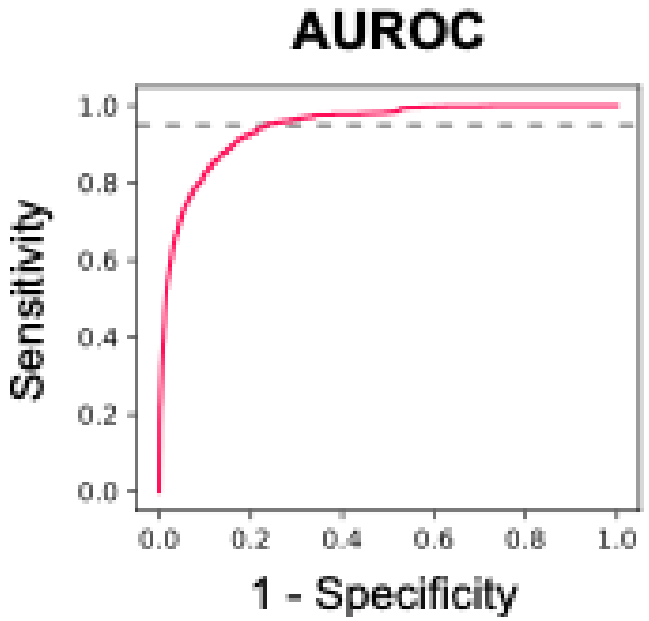
**Predicted risk of transfusion:**

**7%**

**Recommendation:**

**If risk > X%: Order T&S**

**Else: no T&S**



# Evaluation metrics for a surgical transfusion risk model designed to guide T&S order

**AUROC** – probability of rank-ordering 2 patients correctly in likelihood of transfusion

**Sensitivity** – recommending T&S for patients req transfusion

$$\frac{\#\ patients\ predicted\ w\ outcome}{\#\ patients\ w\ outcome}$$

**Specificity** – not recommending T&S for not-transfused

**Positive predictive value** – fraction of patients recommended T&S who are transfused

**# Tests (T&S) recommended** – more straightforward interpretation of positive predictive value and specificity

# Model performance: internal validation

| Model | AUROC | AUPRC | Sensitivity | Specificity | PPV | % rec. T/S |
|---|---|---|---|---|---|---|
| Baseline | 0.888 (0.881-0.894) | 0.215 (0.197-0.235) | 0.970 (0.963-0.977) | 0.439 (0.436-0.442) | 0.037 (0.035-0.038) | 57.0% (56.7-57.3) |
| w/ 5% threshold | | | 0.837 (0.821-0.852) | 0.806 (0.803-0.808) | 0.087 (0.084-0.091) | 20.8% (20.6-21.1) |
| Logistic Regress | 0.907 (0.900-0.913) | 0.280 (0.260-0.300) | 0.962 (0.954-0.970) | 0.512 (0.509-0.515) | 0.042 (0.040-0.043) | 49.8% (49.5-50.1) |
| Decision Tree | 0.916 (0.910-0.922) | 0.298 (0.278-0.318) | 0.963 (0.955-0.971) | 0.618 (0.614-0.621) | 0.053 (0.050-0.055) | 39.5% (39.2-39.8) |
| Random Forest | 0.913 (0.906-0.918) | 0.247 (0.230-0.265) | 0.962 (0.954-0.970) | 0.593 (0.590-0.597) | 0.050 (0.047-0.052) | 41.9% (41.5-42.2) |
| Gradient Boosting Machine | 0.924 (0.919-0.929) | 0.292 (0.273-0.314) | 0.963 (0.956-0.971) | 0.651 (0.648-0.654) | 0.058 (0.055-0.060) | 36.2% (35.9-36.5) |

Lou et al. (2022) *Anesthesiology*.

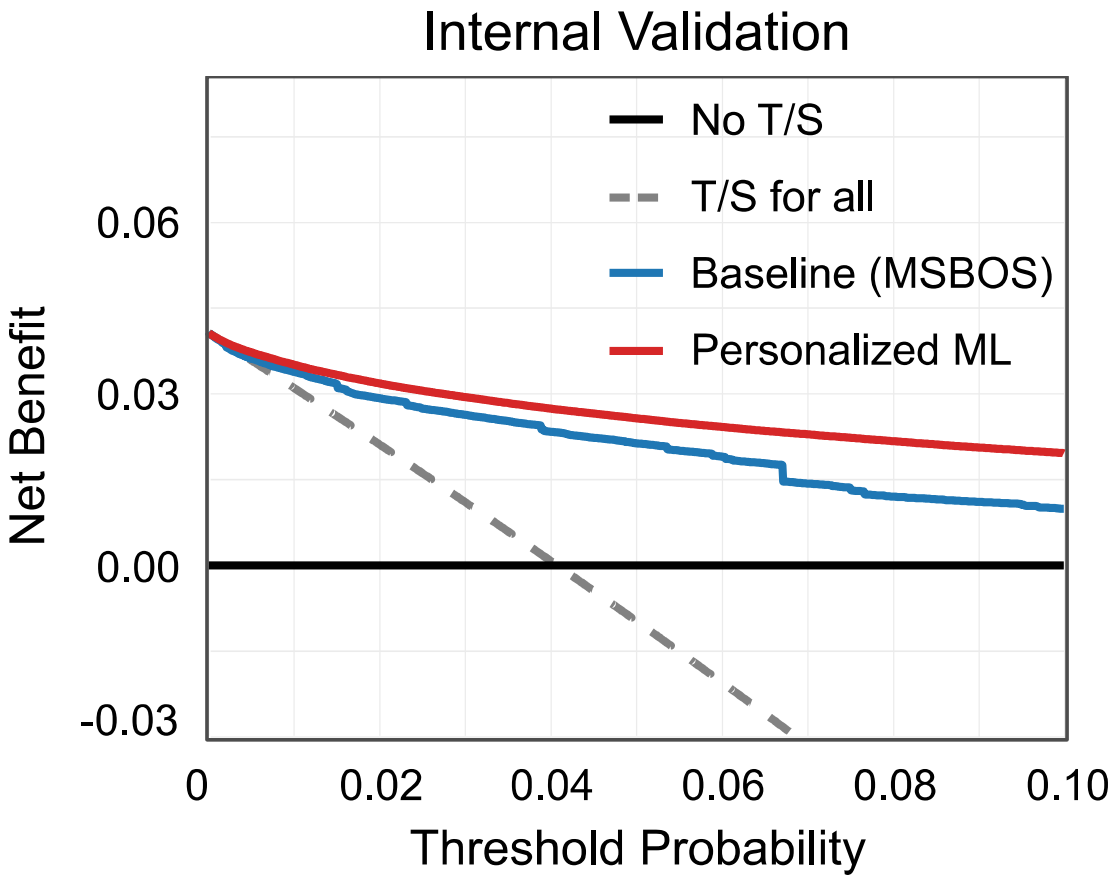
# Personalized model has highest net benefit

Any decision threshold implies a trade-off between false positives and false negatives

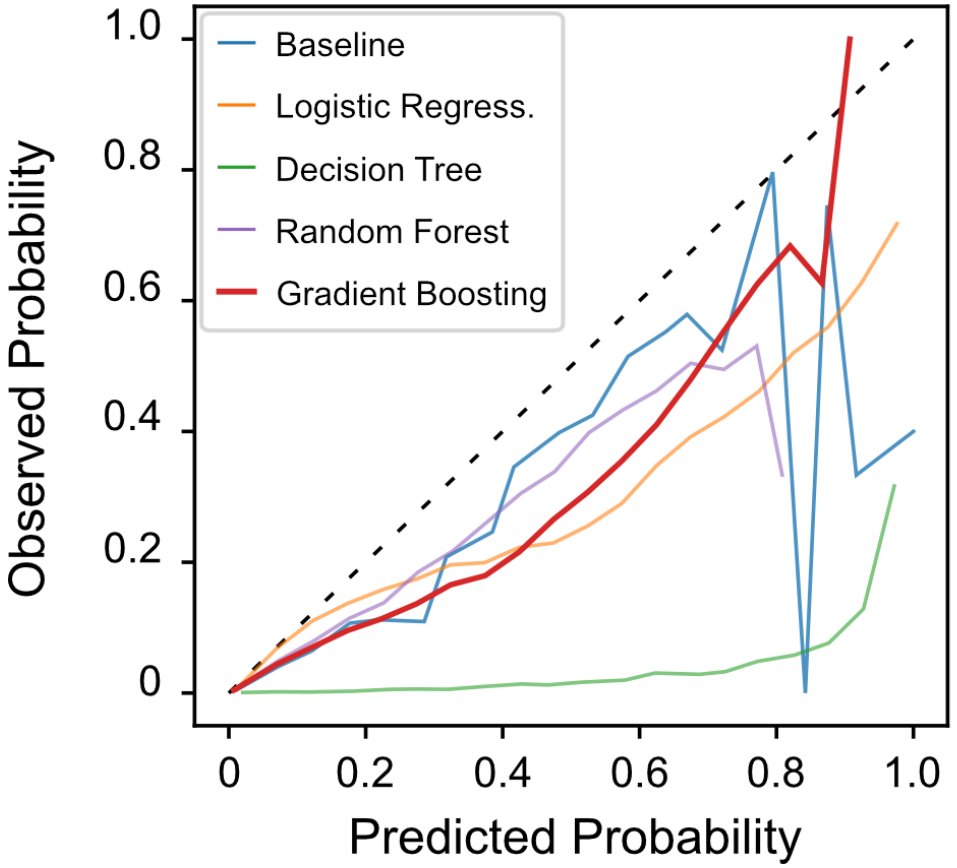
- 5% threshold - willing to perform 20 T/S to identify 1 person who needs transfusion
- 1% threshold - 100:1

Personalized model has highest net benefit at all thresholds > 0.08% (1250:1)

T/S for all patients has highest net benefit < 0.08%



# Model calibration




**Relevant to other model uses**

- Invasive access
  - Large bore intravenous
  - Arterial line
- Surgery location
  - Hospital vs ambulatory surgical center
- Surgical technique
  - Cell salvage

# External validation: BJH/WUSTL

- EHR data is not as clean as NSQIP
  - Patient demographics and labs map 1:1
  - Patient comorbidities are mapped from structured preoperative notes, which are selected based on clinician judgement, rather than strict NSQIP criteria
- Patients are sicker at BJH than NSQIP
- “Transfer learning” – use BJH-specific historical transfusion rates per surgery

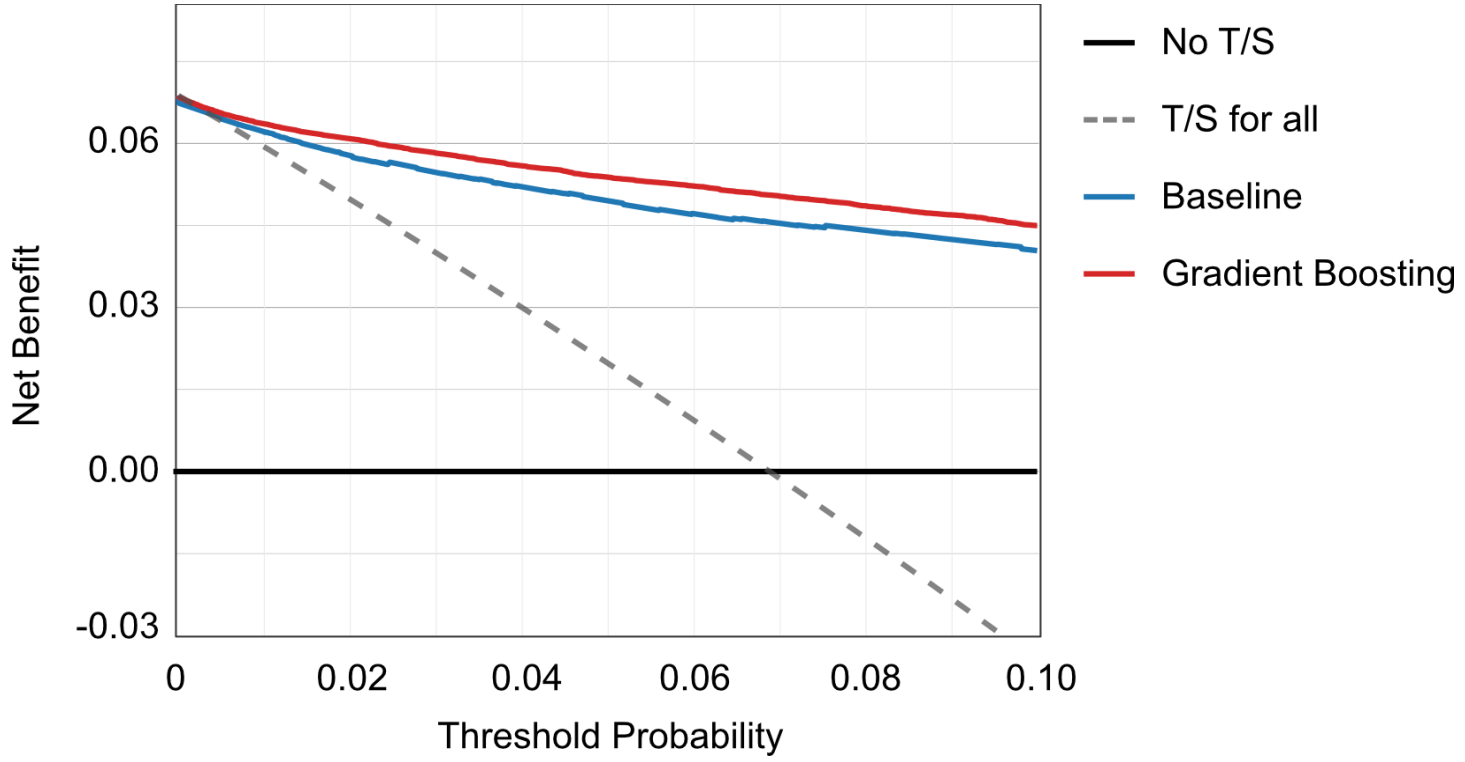
# Model performance at BJH/WUSTL

| Model | AUROC | AUPRC | Sensitivity | Specificity | PPV | % rec T/S | Risk threshold |
|---|---|---|---|---|---|---|---|
| **Baseline** | **0.908** (0.899-0.916) | **0.511** (0.481-0.539) | **0.957** (0.946-0.969) | **0.580** (0.573-0.587) | **0.144** (0.137-0.151) | **45.7%** (45.0-46.4) | **1.6%** |
| w/ 5% threshold | | | **0.911** (0.896-0.927) | **0.727** (0.720-0.733) | **0.198** (0.188-0.207) | **31.7%** (31.1-32.4) | **5.0%** |
| **BJH current** | | | **0.983** | **0.442** | **0.121** | **59.1%** | |
| **Gradient Boosting** | **0.939 (0.933-0.944)** | **0.583 (0.554-0.610)** | **0.959 (0.949-0.969)** | **0.738 (0.732-0.744)** | **0.213 (0.203-0.223)** | **31.0% (30.4-31.6)** | **3.7%** |
| 97% sensitivity | | | **0.971** | **0.692** | **0.189** | **35.4%** | **2.2%** |
| 98% sensitivity | | | **0.982** | **0.566** | **0.143** | **47.2%** | **0.7%** |

# Net benefit at BJC/WUSTL

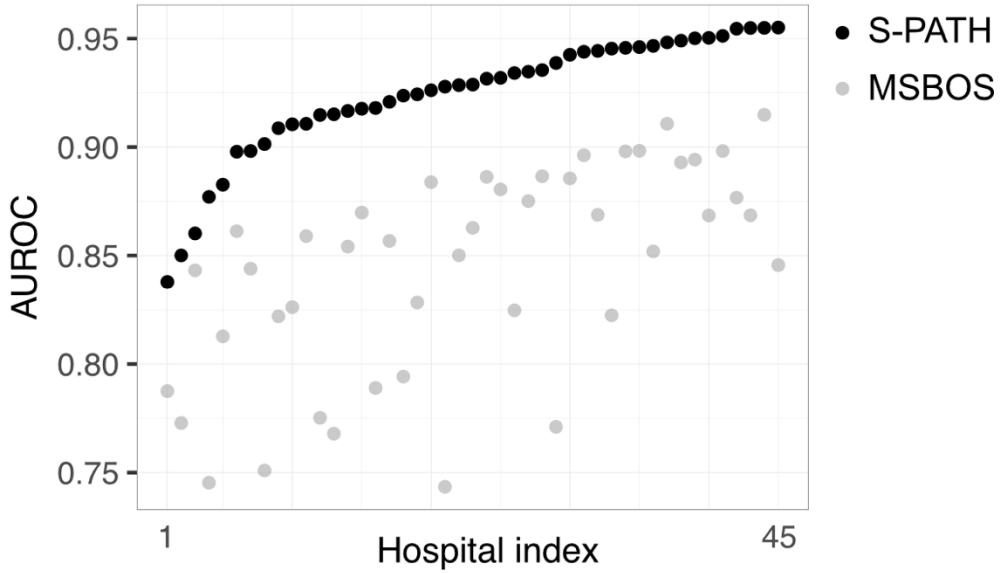
Personalized model has highest net benefit at all thresholds > 0.12% (833:1)

T/S for all patients has highest net benefit < 0.12%



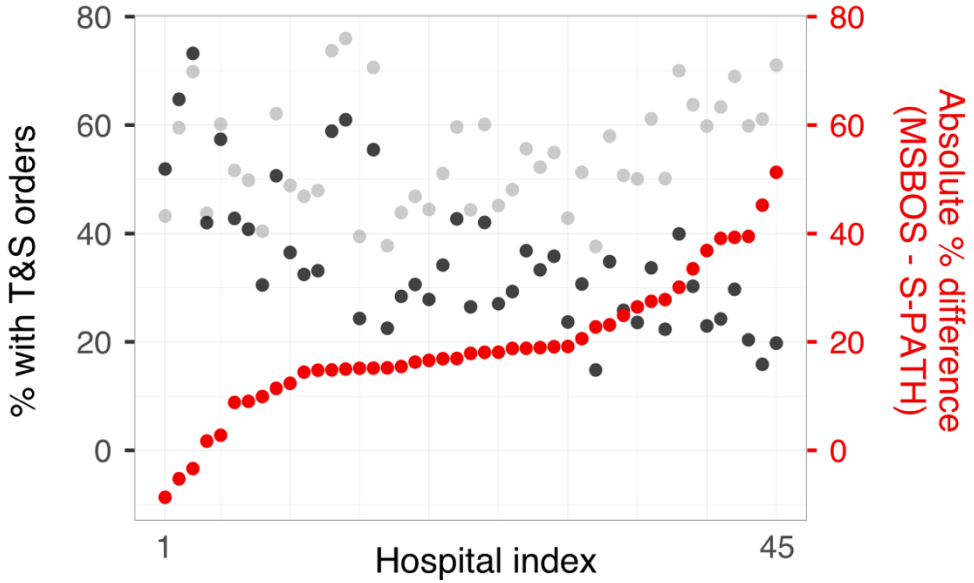
# External validation at 45 hospitals




**Median AUROC**

**0.926** (IQR 0.908-0.955)

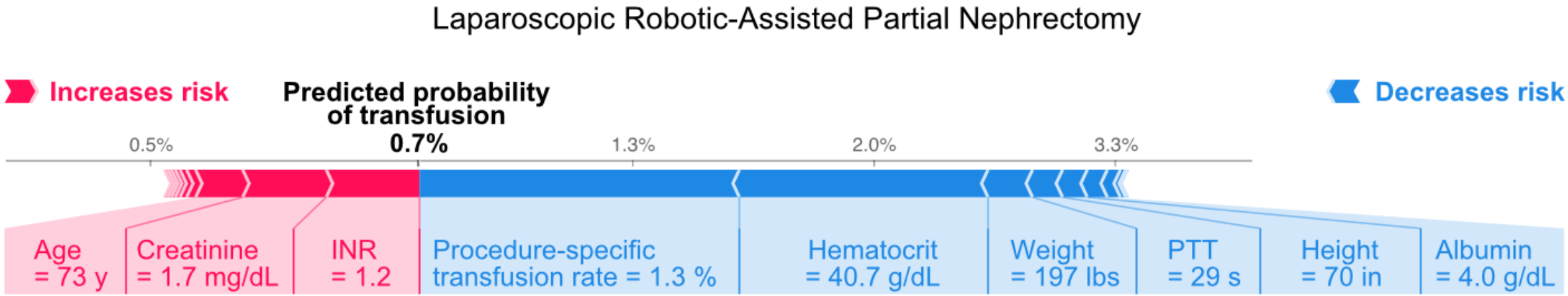



**Median Absolute Difference in % with T&S**

**17.3** (IQR 14.7-26.7)

Lou et al., *JAMA Network Open,* 2025

# Model explanation



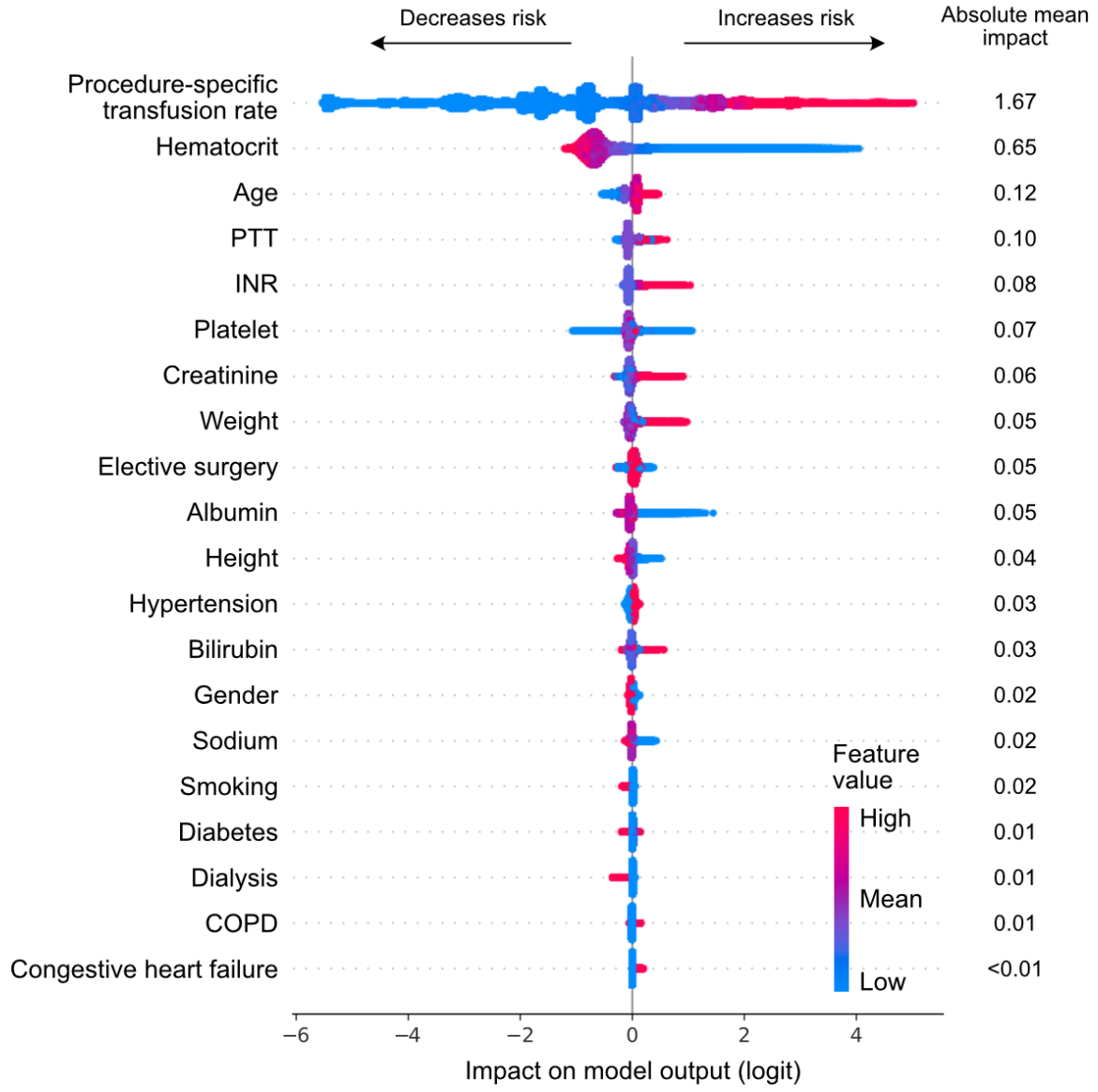
# Model explanation

Color indicates high or low values of each variable

X-axis spread indicates impact on risk

Red to the right of midline indicates high values of that variable increase risk (i.e., INR)

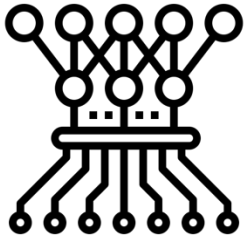
Blue to the right of midline indicates low values of that variable increase risk (i.e., Hct)



# Evaluating the blood prediction model

1. Making the prediction actionable
   - Outcome: Transfusion during surgery
   - Clinical action: Ordering a type and screen
2. Potential biases in the outcome
   - Outcome is objective
   - ? Biases in clinical decision-making ?
3. Training data quality
   - NSQIP is a manually curated, diverse dataset representing >400 hospitals
4. Practical predictors
   - Surgical CPT codes are not often available preoperatively, but alternatives are ok
   - Labs are commonly obtained, but how does timing compare with timing of ordering a type and screen?
5. Clinically meaningful measures of performance
   - Benchmarked against standard of care (MSBOS)

**Clinical input variables**



**Predicted risk of transfusion**

⇩

**Recommendation for T&S**

## S-PATH Model

Robust
- Survived 3 different data extraction pipelines
- Performs well in diverse settings

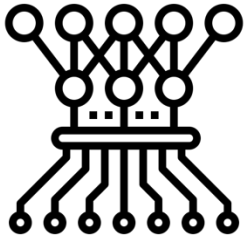
Potential to improve transfusion resource allocation and patient safety

Exists on the internet: bit.ly/spath

Epic

???

**Clinical input variables**



**Predicted risk of transfusion**

⇩

**Recommendation for T&S**

???



## How to collect input data and generate predictions?

Integrate the model within the EHR

## Will clinicians use the tool?

**Human-centered design**

Interviews with key stakeholders to elicit needs, design requirements, workflow

Iterative prototyping to improve design

# Human-centered design

**29 stakeholders**

9 surgeons, 8 anesthesiologists, 4 CRNAs, 4 CPAP NPs, 4 transfusion medicine

## About the Transfusion Risk Score

This Transfusion Risk score uses machine learning to estimate the patient's risk for needing a red blood cell transfusion during their upcoming surgery. **It does NOT take into account the patient's medications, antibody history, or considerations of the surgical approach not included in the procedure name.** Click here for more information on the model.

### Transfusion Risk

- Score calculated 15 hours ago

**0.7%**

| Range | | | Value |
|---|---|---|---|
| 20% | - | 100% | High |
| 3% | - | 20% | Medium |
| 0% | - | 3% | Low |

**Factors Contributing to Score**

| Contribution | Factor | Value |
|---|---|---|
| 32% | % cases transfused | 1.133\|LAPAROSCOPIC TOTAL HYSTERECTOMY |
| 26% | Hematocrit | 41.3 |
| 9% | PTT | 36 |
| 8% | CHF | yes |
| 7% | Age | 38 |
| 7% | INR | 1.04 |
| 2% | Creatinine | 1.28 |
| 2% | Sodium | 142 |

Factors Not Contributing to Score

Description

## Model Recommendation

Consider NOT ordering a type and screen for patients at low risk of transfusion.

Washington University School of Medicine in St. Louis

# Guide to Interpreting the S-PATH Tool

https://sites.wustl.edu/spath-tool/




**1 Risk Score**

Predicted probability that a patient will need a red cell transfusion during surgery.

**2 Risk Range**

To be conservative (i.e., sensitivity 96%), the model recommends type and screens for patients with >3% risk. This is why the risk ranges are labelled with the medium risk band starting at 3%.

RISK RANGE LEGEND

LOW | MEDIUM | HIGH

**3 Factors Contributing to Score**

This box explains how the model arrived at its risk score for this patient. These explanations are customized for each patient and procedure.

- Factor: the variable the model considered
- Contribution: the extent to which the factor contributed to the score for this patient
- Value: the most recent value of the variable
  - Value = -1 means it's missing
  - Value = 1 / 0 means yes / no for binary variables.

| Contribution | Factor | Value |
|---|---|---|
| 40% | Hematocrit | 30 |
| 20% | % cases transfused | 1.4 |

EXAMPLE 1: This patient has a hematocrit of 30%, and that value is 40% of the reason why the model predicted a 5.3% risk. 1.4% of patients having the same procedure have historically needed blood, and that is 20% of the reason why the model predicted 5.3% risk.

| Contribution | Factor | Value |
|---|---|---|
| <1% | Dialysis | 0 |
| <1% | INR | -1 |

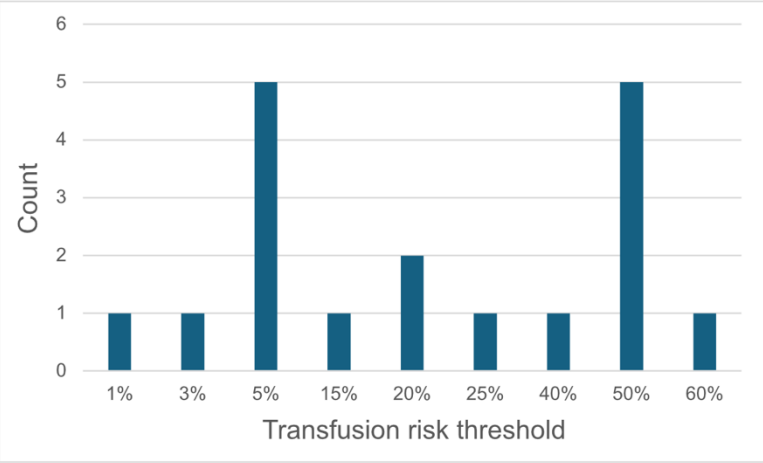
EXAMPLE 2: This patient is not on dialysis (0 = no), and the INR value is missing. These two variables contribute minimally to the model's prediction.

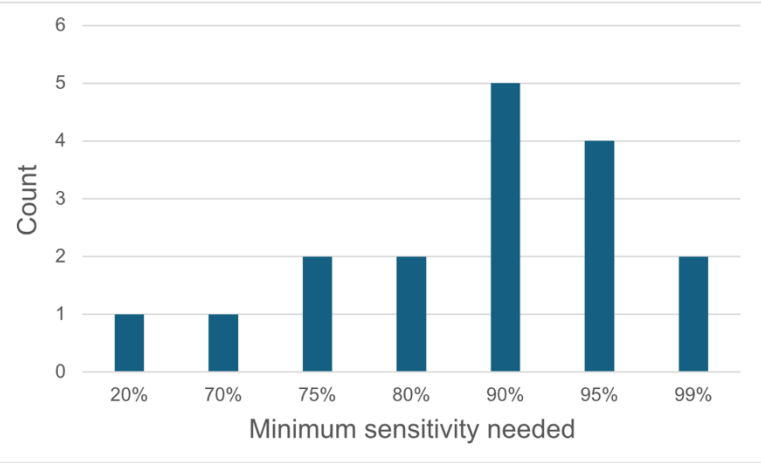
## Known Limitations

- The model does not take into account the patient's preoperative medications, red cell alloantibody history, or considerations of the surgical approach not included in the procedure name (for example, anatomic proximity to major blood vessels).
- Due to EHR limitations, the model may be misinformed about certain risk factors. For example, patient hematocrit values in CareEverywhere are not taken into account. If the primary procedure name is not an accurate representation of the planned surgery, the tool's risk estimate may be inaccurate as well.
- The model only makes predictions on the potential need for red blood cell transfusion. It does not advise on the potential need for platelets or other clotting factors.

# Design requirements (N = 20)

Recommendation threshold (i.e., > **X%** = T&S recommended)




**Final Decision**

**3%** risk threshold

**96%** sensitivity

Simple display within existing navigator, minimal clicks

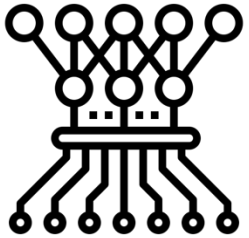
Additional information hidden by default but expandable

Yang et al. (2024) *Transfusion*, doi: 10.1111/trf.18012

Epic

???

**Clinical input variables**



**Predicted risk of transfusion**

⇩

**Recommendation for T&S**

???



# How to collect input data and generate predictions?

Integrate the model within the EHR

Cognitive Computing



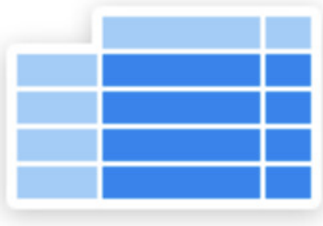
Trigger



Reporting workbench

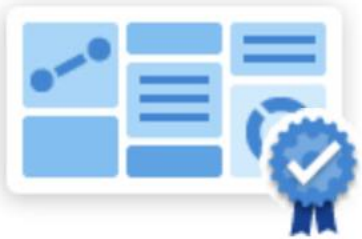
NebulaCloud





Flowsheet Row





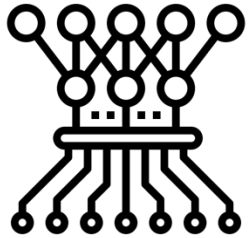
Print Group



Epic

???

**Clinical input variables**



**Predicted risk of transfusion**

⇩

**Recommendation for T&S**

???



# How to collect input data and generate predictions?

## Integrate the model within the EHR

Cognitive Computing

### About the Transfusion Risk Score

This Transfusion Risk score uses machine learning to estimate the patient's risk for needing a red blood cell transfusion during their upcoming surgery. **It does NOT take into account the patient's medications, antibody history, or considerations of the surgical approach not included in the procedure name.** Click here for more information on the model.

**Transfusion Risk**

- Score calculated 15 hours ago

**0.7%**

| Range | | | Value |
|---|---|---|---|
| 20% | - | 100% | High |
| 3% | - | 20% | Medium |
| 0% | - | 3% | Low |

**Factors Contributing to Score**

| Contribution | Factor | Value |
|---|---|---|
| 32% | % cases transfused | 1.133\|LAPAROSCOPIC TOTAL HYSTERECTOMY |
| 26% | Hematocrit | 41.3 |
| 9% | PTT | 36 |
| 8% | CHF | yes |
| 7% | Age | 38 |
| 7% | INR | 1.04 |
| 2% | Creatinine | 1.28 |
| 2% | Sodium | 142 |

**Factors Not Contributing to Score**

**Description**

### Model Recommendation

Consider NOT ordering a type and screen for patients at low risk of transfusion.

# Prospective validation (Nov ’24 – Jul ‘25)

## Efficiency Evaluation

**7454 / 10278 (72%)** patients seen by CPAP had SPATH scores

- Walk-in patients and rare procedures won’t have scores

**S-PATH T&S: 3528 (47%)** **Usual care T&S: 5981 (80%)**

- Absolute reduction of 33% (2453 *$400 ~ $980k)

- Differences due to pts having low risk surgery and nl Hct

# Prospective validation (Nov ’24 – Jul ‘25)

## Patient Safety Evaluation

**609 (9%)** patients required RBC transfusion during surgery

- **Usual care missed 116 (19%)**: no T&S by start of case
  - **31 (5%)** had no T&S by 1h after start of case
  - All of these patients were recommended to have T&S by S-PATH
- **S-PATH missed 26 (4%)** = 96% sensitivity
  - Most reflect that unexpected bleeding can occur during surgery
  - Some cases where it was not obvious that transfusion was indicated

# Summary

Developed and validated S-PATH as a surgical transfusion risk prediction model, designed to guide T&S decisions

Designed the user interface to be easy to interpret and well-integrated with clinical workflow

Integrated S-PATH within the EHR and prospectively validated performance

Potential to improve patient safety and resource allocation if used as clinical decision support

# Proposed cluster-randomized trial

**Setting:** Preop Assessment Clinic (CPAP)

**Participants**

- NPs, intern physicians -> randomized to have access to S-PATH tool
- Patients seen in person (not TPAP), randomized based on clinician assigned to them, requesting IRB waiver of informed consent
- Patients excluded if red cell alloantibodies, surgery not in main ORs at BJH

**Stepped-wedge design**

Wash-in period to test technology and training

