Tips for clinically meaningful machine learning in healthcare

Sunny Lou, MD PhD
2023.10.05
Great interest in ML for health

The AI Doctor Will See You Now
Artificial intelligence comes to the doctor’s office, helping identify disease, monitor heart activity, stave off seizures

221,931 results!!

2000

2022

52,000

Washington University School of Medicine in St. Louis

Department of Anesthesiology
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, BS1; Dennis L. Shung, MD, PhD2; Alyssa A. Grimshaw, MSLIS3; 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
<table>
<thead>
<tr>
<th>Not fit for purpose</th>
<th>No validation</th>
<th>No implementation</th>
<th>Not adopted</th>
</tr>
</thead>
<tbody>
<tr>
<td>Developed on wrong patient population</td>
<td>Lack of data or incentive to pursue validation studies</td>
<td>No impact on decision making or patient (health) outcomes</td>
<td>Prediction (perceived as) not useful</td>
</tr>
<tr>
<td>Expensive or non-available predictors</td>
<td>Incompletely reported prediction model</td>
<td>No software developed to implement and use the model</td>
<td>Predictions not trusted</td>
</tr>
<tr>
<td>Time intensive to use model</td>
<td>Poorly developed or overfitted model</td>
<td>Requirements for adherence to (medical device) regulations</td>
<td>Model not transparent enough, or no tools available to enhance its use in practice</td>
</tr>
<tr>
<td>Outcome measured unreliably</td>
<td>Proprietary model code</td>
<td>Cost-effectiveness of using proprietary model</td>
<td>Model (perceived as) outdated</td>
</tr>
</tbody>
</table>

**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?
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?
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
Two stories on ML in healthcare

Clinical decision support for perioperative blood management

Risk stratification for physician burnout
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

<table>
<thead>
<tr>
<th>Service</th>
<th>Procedure</th>
<th>Routine blood bank order</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cardiac Surgery</td>
<td>Pacemaker/AICD insertion (without extraction)</td>
<td>Type and Screen</td>
</tr>
<tr>
<td></td>
<td>Pacemaker/AICD removal</td>
<td>Type and Cross (2)</td>
</tr>
<tr>
<td></td>
<td>Coronary Bypass and Open Valve Cases</td>
<td>Type and Cross (4)</td>
</tr>
<tr>
<td></td>
<td>Thoracic Aortic Surgery</td>
<td>Type and Cross (4)</td>
</tr>
<tr>
<td></td>
<td>Transcatheter Aortic Valve Surgery (TAVR)</td>
<td>Type and Cross (4)</td>
</tr>
<tr>
<td></td>
<td>ECMO</td>
<td>Type and Cross (6)</td>
</tr>
<tr>
<td></td>
<td>Lung transplant surgery</td>
<td>Type and Cross (6)</td>
</tr>
<tr>
<td></td>
<td>Redo Sternotomy for Heart Surgery</td>
<td>Type and Cross (6)</td>
</tr>
<tr>
<td></td>
<td>Ventricular Assist Device (VAD)</td>
<td>Type and Cross (6)</td>
</tr>
<tr>
<td></td>
<td>Heart transplant surgery</td>
<td>Type and Cross (10)</td>
</tr>
<tr>
<td></td>
<td>Transcatheter Mitral Valve Repair</td>
<td>Type and Screen</td>
</tr>
</tbody>
</table>

**Hospital-specific**

<table>
<thead>
<tr>
<th>Service</th>
<th>Procedure</th>
<th>Routine blood bank order</th>
</tr>
</thead>
<tbody>
<tr>
<td>Electrophysiology</td>
<td>Pacemaker/AICD Insertion or Revision</td>
<td>None (Emergency Release Blood)</td>
</tr>
<tr>
<td></td>
<td>Cardiac Ablation</td>
<td>None (Emergency Release Blood)</td>
</tr>
<tr>
<td></td>
<td>Watchman device/Lariat Procedure (left atrial appendage exclusion)</td>
<td>Type and Cross (2)</td>
</tr>
</tbody>
</table>

**NOT patient-specific**

<table>
<thead>
<tr>
<th>Service</th>
<th>Procedure</th>
<th>Routine blood bank order</th>
</tr>
</thead>
<tbody>
<tr>
<td>ENT</td>
<td>Facial Fractures</td>
<td>None (Emergency Release Blood)</td>
</tr>
<tr>
<td></td>
<td>Laryngoscopy with Biopsy</td>
<td>None (Emergency Release Blood)</td>
</tr>
<tr>
<td></td>
<td>Thyroidectomy/Parathyroidectomy</td>
<td>None (Emergency Release Blood)</td>
</tr>
<tr>
<td></td>
<td>Tonsillectomy and/or Adenoidectomy</td>
<td>None (Emergency Release Blood)</td>
</tr>
<tr>
<td></td>
<td>Hemiglossectomy</td>
<td>Type and Screen</td>
</tr>
</tbody>
</table>
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 PUF

- 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

<table>
<thead>
<tr>
<th>Variable</th>
<th>Data Type</th>
<th>Training: (n = 3,049,617)</th>
<th>Internal Validation: (n = 1,076,441)</th>
<th>External Validation: (n = 16,053)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transfused – Yes, n (%)</td>
<td>Binary</td>
<td>73,313 (2.4%)</td>
<td>23,205 (2.2%)*</td>
<td>1,104 (6.7%)*</td>
</tr>
<tr>
<td>Age, mean (std)</td>
<td>Continuous</td>
<td>57 (17)</td>
<td>57 (17)*</td>
<td>57 (16)</td>
</tr>
<tr>
<td>Height (in), mean (std)</td>
<td>Continuous</td>
<td>66 (4)</td>
<td>66 (4)*</td>
<td>67 (4)*</td>
</tr>
<tr>
<td>Weight (lb), mean (std)</td>
<td>Continuous</td>
<td>188 (51)</td>
<td>186 (49)*</td>
<td>193 (56)*</td>
</tr>
<tr>
<td>Gender – Male, n (%)</td>
<td>Binary</td>
<td>1,316,142 (43%</td>
<td>454,517 (42%)*</td>
<td>5,962 (47%)*</td>
</tr>
<tr>
<td>ASA Status – I, n (%)</td>
<td>Ordinal</td>
<td>253,948 (8%)</td>
<td>57,952 (4%)</td>
<td>7,592 (4%)</td>
</tr>
<tr>
<td>ASA Status – II, n (%)</td>
<td></td>
<td>1,356,498 (45%)</td>
<td>486,469 (45%)*</td>
<td>1,142 (11%)*</td>
</tr>
<tr>
<td>ASA Status – III, n (%)</td>
<td></td>
<td>1,247,668 (41%)</td>
<td>442,515 (41%)*</td>
<td>7,504 (47%)*</td>
</tr>
<tr>
<td>ASA Status – IV, n (%)</td>
<td></td>
<td>178,613 (6%)</td>
<td>57,950 (5%)</td>
<td>1,881 (12%)*</td>
</tr>
<tr>
<td>ASA Status – V, n (%)</td>
<td></td>
<td>5,413 (0.2%)</td>
<td>1,905 (0.2%)</td>
<td>69 (0.4%)*</td>
</tr>
<tr>
<td>HTN – Yes, n (%)</td>
<td>Binary</td>
<td>1,354,643 (44%)</td>
<td>466,520 (43%)*</td>
<td>7,227 (45%)</td>
</tr>
<tr>
<td>CHF – Yes, n (%)</td>
<td>Binary</td>
<td>25,930 (1%)</td>
<td>9,051 (1%)</td>
<td>1,192 (7%)</td>
</tr>
<tr>
<td>Smoking – Yes, n (%)</td>
<td>Binary</td>
<td>516,646 (17%)</td>
<td>165,553 (15%)*</td>
<td>1,760 (11%)*</td>
</tr>
<tr>
<td>COPD – Yes, n (%)</td>
<td>Binary</td>
<td>129,528 (4%)</td>
<td>43,290 (4%)*</td>
<td>1,137 (8%)*</td>
</tr>
<tr>
<td>Dialysis – Yes, n (%)</td>
<td>Binary</td>
<td>39,343 (1%)</td>
<td>11,459 (1%)</td>
<td>524 (3%)*</td>
</tr>
<tr>
<td>Diabetes – No, n (%)</td>
<td>Ordinal</td>
<td>2,574,255 (8%)</td>
<td>913,163 (8%)</td>
<td>13,192 (8%)*</td>
</tr>
<tr>
<td>Meds, n (%)</td>
<td></td>
<td>300,959 (10%)</td>
<td>106,689 (10%)*</td>
<td>1,881 (12%)*</td>
</tr>
<tr>
<td>Insulin, n (%)</td>
<td></td>
<td>174,402 (6%)</td>
<td>56,589 (5%)*</td>
<td>979 (6%)*</td>
</tr>
</tbody>
</table>

* indicates Bonferroni corrected p < 0.05 compared to training data
Model evaluation

Discrimination
- AUROC (c-statistic)
- AUPRC (average precision)

Calibration
- Calibration plot

Utility
- Net benefit

For the T/S use case:
- Two thresholds:
  - 5%, recommended by guidelines
  - To achieve 96% sensitivity
- Metrics
  - Positive predictive value
  - # of T/S recommended (%)
## Model performance: internal validation

<table>
<thead>
<tr>
<th>Model</th>
<th>AUROC</th>
<th>AUPRC</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>PPV</th>
<th>% rec. T/S</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.888</td>
<td>0.215</td>
<td>0.970</td>
<td>0.439</td>
<td>0.037</td>
<td>57.0%</td>
</tr>
<tr>
<td>w/ 5% threshold</td>
<td>0.837</td>
<td>0.298</td>
<td>0.806</td>
<td>0.087</td>
<td>0.053</td>
<td>20.8%</td>
</tr>
<tr>
<td>Logistic Regress</td>
<td>0.907</td>
<td>0.280</td>
<td>0.962</td>
<td>0.512</td>
<td>0.042</td>
<td>49.8%</td>
</tr>
<tr>
<td>Decision Tree</td>
<td>0.916</td>
<td>0.298</td>
<td>0.963</td>
<td>0.618</td>
<td>0.053</td>
<td>39.5%</td>
</tr>
<tr>
<td>Random Forest</td>
<td>0.913</td>
<td>0.247</td>
<td>0.962</td>
<td>0.593</td>
<td>0.050</td>
<td>41.9%</td>
</tr>
<tr>
<td>Gradient Boosting Machine</td>
<td>0.924</td>
<td>0.292</td>
<td>0.963</td>
<td>0.651</td>
<td>0.058</td>
<td>36.2%</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Model</th>
<th>AUROC</th>
<th>AUPRC</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>PPV</th>
<th>% rec T/S</th>
<th>Risk threshold</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Baseline</strong> w/ 5% threshold</td>
<td>0.908</td>
<td>0.511</td>
<td>0.957</td>
<td>0.580</td>
<td>0.144</td>
<td>45.7%</td>
<td>1.6%</td>
</tr>
<tr>
<td></td>
<td>(0.899-0.916)</td>
<td>(0.481-0.539)</td>
<td>(0.946-0.969)</td>
<td>(0.573-0.587)</td>
<td>(0.137-0.151)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>BJH current</strong></td>
<td>0.911</td>
<td>0.727</td>
<td>0.983</td>
<td>0.442</td>
<td>0.121</td>
<td>59.1%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.896-0.927)</td>
<td>(0.720-0.733)</td>
<td></td>
<td>(0.188-0.207)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Gradient Boosting</strong></td>
<td>0.939</td>
<td>0.583</td>
<td>0.959</td>
<td>0.738</td>
<td>0.213</td>
<td>31.7%</td>
<td>5.0%</td>
</tr>
<tr>
<td></td>
<td>(0.933-0.944)</td>
<td>(0.554-0.610)</td>
<td>(0.949-0.969)</td>
<td>(0.732-0.744)</td>
<td>(0.203-0.223)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>97% sensitivity</td>
<td>0.971</td>
<td>0.692</td>
<td>0.189</td>
<td>35.4%</td>
<td>2.2%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>98% sensitivity</td>
<td>0.982</td>
<td>0.566</td>
<td>0.143</td>
<td>47.2%</td>
<td>0.7%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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%
Model explanation

Laparoscopic Robotic-Assisted Partial Nephrectomy

- **Increases risk**
  - Predicted probability of transfusion: 0.7%
  - Age: 73 y
  - Creatinine: 1.7 mg/dL
  - INR: 1.2
  - Procedure-specific transfusion rate: 1.3%

- **Decreases risk**
  - Hematocrit: 40.7 g/dL
  - Weight: 197 lbs
  - PTT: 29 s
  - Height: 70 in
  - Albumin: 4.0 g/dL
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)
Implementation at WashU

## Needs Assessment
- Semi-structured interviews
  - 20 stakeholders
  - Interview guide based on Normalization Process Theory

## Iterative Prototyping
- Design workshops
  - Update design based on needs, identified issues
- Scenario Evaluations
  - Users try out design, identify usability issues

### Transfusion Risk Score

**About the Transfusion Risk Score**

This Transfusion Risk score uses machine learning to estimate this 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](#).

**Surgical Transfusion Risk**

Banks, George A - Score calculated 15 minutes ago

![Transfusion Risk Score](image)

#### Factors Contributing to Score

<table>
<thead>
<tr>
<th>Factor</th>
<th>Value</th>
<th>Factor</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hematocrit</td>
<td>30</td>
<td>Sodium</td>
<td>130</td>
</tr>
<tr>
<td>% cases transfused</td>
<td>1.4</td>
<td>Weight (lbs)</td>
<td>198</td>
</tr>
<tr>
<td>Age</td>
<td>78</td>
<td>INR</td>
<td>-1</td>
</tr>
<tr>
<td>Platelet count</td>
<td>125</td>
<td>CO2</td>
<td>0</td>
</tr>
<tr>
<td>Platelet count</td>
<td>125</td>
<td>HTN</td>
<td>1</td>
</tr>
<tr>
<td>Sex</td>
<td>male</td>
<td>PTT</td>
<td>-1</td>
</tr>
<tr>
<td>Height (in)</td>
<td>67</td>
<td>Bilirubin</td>
<td>-1</td>
</tr>
<tr>
<td>CHF</td>
<td>1</td>
<td>Albumin</td>
<td>-1</td>
</tr>
<tr>
<td>Creatinine</td>
<td>1.4</td>
<td>Elective surgery</td>
<td>0</td>
</tr>
</tbody>
</table>

**Model Recommendation**

Consider ordering a type and screen for patients at **medium** or **high** risk of transfusion. These risk thresholds are set to be conservative, so you can expect the model to recommend type and screens for 96 out of every 100 patients who need blood during surgery.

Next steps
- User evaluation study
- Stakeholder buy-in
- Cluster-randomized trial
Evaluating the blood prediction project

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)
Physician burnout

- Emotional exhaustion
  i.e. lack of enthusiasm

- Depersonalization
  i.e. lack of empathy

- Professional Accomplishment
  i.e. lack of meaning

50% of practicing physicians

70% of trainee physicians (residents, fellows)
Consequences of burnout

- Substance Abuse
- Car accidents
- Suicide

- Medical error
- Poor care quality
- Poor satisfaction

- Increased turnover
- Decreased work hours
- > $4 billion dollars
Measurement of burnout is challenging

Current status

Survey data

Cross-sectional

What if we could screen for burnout using EHR audit logs?
Clinical work occurs overwhelmingly within the EHR

- Orders
- Documentation
- Communication

50% of time at work is spent using the EHR

100% of clinical care passes through the EHR
Audit logs are a fingerprint for clinicians’ work habits

Mandated for security reasons
Captures whenever patient data is viewed or modified
Tracks who performed what on which patient’s chart

Can we use audit logs to measure work-related behaviors?

<table>
<thead>
<tr>
<th>ACCESS_TIME</th>
<th>USER_ID</th>
<th>PAT_ID</th>
<th>ACTION_PERFORMED</th>
</tr>
</thead>
<tbody>
<tr>
<td>12/21 7:20:17</td>
<td>Z</td>
<td></td>
<td>Inpatient Patient Lists list loaded</td>
</tr>
<tr>
<td>12/21 7:20:17</td>
<td>Z</td>
<td></td>
<td>Inpatient system list accessed</td>
</tr>
<tr>
<td>12/21 7:20:22</td>
<td>Z</td>
<td>A</td>
<td>Results Review accessed</td>
</tr>
<tr>
<td>12/21 7:20:32</td>
<td>Z</td>
<td>A</td>
<td>Report with patient data viewed</td>
</tr>
<tr>
<td>12/21 7:20:32</td>
<td>Z</td>
<td>A</td>
<td>Report with patient data viewed</td>
</tr>
<tr>
<td>12/21 7:20:34</td>
<td>Z</td>
<td>A</td>
<td>Imaging PACS accessed</td>
</tr>
<tr>
<td>12/21 7:21:25</td>
<td>Z</td>
<td></td>
<td>Inpatient system list accessed</td>
</tr>
<tr>
<td>12/21 7:21:29</td>
<td>Z</td>
<td>B</td>
<td>Storyboard viewed</td>
</tr>
<tr>
<td>12/21 7:21:29</td>
<td>Z</td>
<td>B</td>
<td>Visit Navigator template loaded</td>
</tr>
<tr>
<td>12/21 7:21:30</td>
<td>Z</td>
<td>B</td>
<td>Orders section accessed</td>
</tr>
<tr>
<td>12/21 7:21:30</td>
<td>Z</td>
<td>B</td>
<td>Order sets accessed</td>
</tr>
<tr>
<td>12/21 7:22:02</td>
<td>Z</td>
<td>B</td>
<td>Order list changed</td>
</tr>
<tr>
<td>12/21 7:22:07</td>
<td>Z</td>
<td>B</td>
<td>Chart Review Notes tab selected</td>
</tr>
<tr>
<td>12/21 7:22:09</td>
<td>Z</td>
<td>B</td>
<td>Chart Review Note report viewed</td>
</tr>
<tr>
<td>12/21 7:28:44</td>
<td>Z</td>
<td>B</td>
<td>Chart Review Note report viewed</td>
</tr>
</tbody>
</table>
Digital Phenotyping for Burnout

Digital Phenotyping for Burnout

**Workload features**
- Patient load
- Total EHR time
- After-hours time
- Inbox time
- Chart review time
- Note time
- Order volume

**Temporal features**
Statistical features (i.e., mean, min, max, skew, kurtosis, energy, entropy, autocorrelation, slope):
- Temporal pattern of EHR click activities
- Temporal pattern of attention switching
Study design and data collection

EHR-based workload is associated with burnout

Multivariable mixed-effect model with random intercept per participant, and fixed effects controlling for specialty and gender. Dot shows the estimated effect of a 25th to 75th percentile change in each variable (with shaded area showing 95% CI)

Machine learning evaluation plan

528 surveys sent to 88 participants
- 112 surveys not completed
- 416 surveys completed by 88 participants
- 25 surveys excluded due to lack of audit log data
- 391 surveys with associated 10,045,218 EHR audit log actions analyzed from 88 participants

Nested cross-validation (Grouped by participants)
- 20x
- Average model performance across all 200 repeated evaluations

Outer 10-fold CV Loop
- Testing Set
- Training Set

Inner 10-fold CV Loop
- Normalization
- Best hyperparameters
- Inner 10-fold CV used to identify optimal hyperparameters: cross-validated random search over parameter settings to maximize average AUROC over 10 folds

10 evaluation score sets (AUROC & accuracy)

Lou et al. (2022) Journal of Biomedical Informatics.
Workload and temporal features weakly predict burnout

<table>
<thead>
<tr>
<th>Feature Set</th>
<th>Best Model</th>
<th>Mean Absolute Error</th>
<th>AUROC</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Workload</td>
<td>Random Forest</td>
<td>0.602 (0.412, 0.826)</td>
<td>0.595 (0.355, 0.808)</td>
<td>0.567 (0.393, 0.742)</td>
</tr>
<tr>
<td>Temporal</td>
<td>Support Vector Machine</td>
<td>0.596 (0.391, 0.826)</td>
<td>0.581 (0.343, 0.790)</td>
<td>0.556 (0.318, 0.756)</td>
</tr>
<tr>
<td>Workload + Temporal</td>
<td>Gradient Boosting Machine</td>
<td>0.619 (0.438, 0.844)</td>
<td>0.583 (0.270, 0.831)</td>
<td>0.559 (0.386, 0.780)</td>
</tr>
</tbody>
</table>

Median PFI burnout score: 1.2 (IQR 0.7-1.7)
PFI score > 1.33 used to indicate burnout

Lou et al. (2022) Journal of Biomedical Informatics.
Baseline burnout is highly predictive of future burnout

<table>
<thead>
<tr>
<th>Feature Set</th>
<th>Best Model</th>
<th>Mean Absolute Error</th>
<th>AUROC</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Workload</td>
<td>Random Forest</td>
<td>0.602 (0.412, 0.826)</td>
<td>0.595 (0.355, 0.808)</td>
<td>0.567 (0.393, 0.742)</td>
</tr>
<tr>
<td>Temporal</td>
<td>Support Vector Machine</td>
<td>0.596 (0.391, 0.826)</td>
<td>0.581 (0.343, 0.790)</td>
<td>0.556 (0.318, 0.756)</td>
</tr>
<tr>
<td>Workload + Temporal</td>
<td>Gradient Boosting Machine</td>
<td>0.619 (0.438, 0.844)</td>
<td>0.583 (0.270, 0.831)</td>
<td>0.559 (0.386, 0.780)</td>
</tr>
<tr>
<td>First Survey Score + Workload</td>
<td>Neural Network</td>
<td>0.423 (0.293, 0.567)</td>
<td>0.829 (0.607, 0.996)</td>
<td>0.781 (0.587, 0.936)</td>
</tr>
<tr>
<td>First Survey Score</td>
<td>Neural Network</td>
<td>0.432 (0.304, 0.570)</td>
<td>0.819 (0.551, 0.999)</td>
<td>0.765 (0.547, 0.952)</td>
</tr>
</tbody>
</table>

Lou et al. (2022) Journal of Biomedical Informatics.
Lessons Learned

Resident physicians may not be the right population to develop a stable model – not “steady state”

Digital phenotyping for burnout from EHR logs is hard
- Inter-individual variability in response to workload
- Inability to capture non-EHR work, personal, environmental factors that also contribute to burnout
Evaluating the burnout project

1. Making the prediction actionable
   - Outcome: Burnout as defined by PFI survey
   - Clinical action: Interventions are unclear (scheduling, training, resources)

2. Potential biases in the outcome
   - Outcome is subjective, has reasonable validity but is imperfect

3. Training data quality
   - Small dataset (~80 people, ~300 surveys), maybe wrong population, not representative
   - Audit logs are reproducible and automatically generated

4. Practical predictors
   - Audit logs are required by law but are data intensive

5. Clinically meaningful measures of performance
   - AUROC because clinical action was unclear
   - Didn’t measure calibration

Department of Anesthesiology
Washington University School of Medicine in St. Louis
Thanks!

Sunny Lou, slou@wustl.edu
Thomas Kannampallil
Michael Avidan
Sachin Kheterpal
Chenyang Lu
Hanyang Liu
Ben Warner
Bruce Hall
Stephen Gregory

Troy Wildes
Ron Jackups
Joanna Abraham
Jing Li
Jennifer Duncan
Brad Evanoff
Sayantan Kumar
Derek Harford
Daphne Lew