Predicting Physician Burnout Using EHR Activity Logs

Hanyang Liu, PhD Student
Advisor: Chenyang Lu
Oct 12, 2023
Background
Physician Burnout

**Burnout**
-- “mental exhaustion caused by one’s professional life”

**Consequence**
- Substance Abuse
- Suicidal Ideation
- Medical error
- Poor care quality
- Increased turnover
- Decreased work hours
- > $4 billion /year

50% of physicians

70% of trainee physicians (residents, fellows)

COVID-19 pandemic further fueled the negative impact
Measurement and Intervention

Traditional Burnout Assessment
- Physician → Survey → Burnout Assessment
- Obtrusive, Slow and Expensive

Personalized Intervention
- Requires monitoring & prediction system
- No such system yet!
Approach I: Feature Engineering + Standard ML
### What are EHR Activity Logs?

Tracks **who** performed **what** at **what time**.

<table>
<thead>
<tr>
<th>Time</th>
<th>USER_ID</th>
<th>PAT_ID</th>
<th>Clinical Action</th>
</tr>
</thead>
<tbody>
<tr>
<td>12/21/19 7:21</td>
<td>M58530</td>
<td></td>
<td>Inpatient Patient Lists list loaded</td>
</tr>
<tr>
<td>12/21/19 7:21</td>
<td>M58530</td>
<td></td>
<td>Inpatient system list accessed</td>
</tr>
<tr>
<td>12/21/19 7:21</td>
<td>M58530</td>
<td>A</td>
<td>Storyboard viewed</td>
</tr>
<tr>
<td>12/21/19 7:21</td>
<td>M58530</td>
<td>A</td>
<td>SmartSets activity selected</td>
</tr>
<tr>
<td>12/21/19 7:21</td>
<td>M58530</td>
<td>A</td>
<td>Visit Navigator template loaded</td>
</tr>
<tr>
<td>12/21/19 7:21</td>
<td>M58530</td>
<td>A</td>
<td>Problem List accessed</td>
</tr>
<tr>
<td>12/21/19 7:21</td>
<td>M58530</td>
<td>A</td>
<td>Chart Review Encounters tab selected</td>
</tr>
<tr>
<td>12/21/19 7:21</td>
<td>M58530</td>
<td>A</td>
<td>Chart Review Notes tab selected</td>
</tr>
<tr>
<td>12/21/19 7:21</td>
<td>M58530</td>
<td>A</td>
<td>Chart Review Note tab selected</td>
</tr>
<tr>
<td>12/21/19 7:22</td>
<td>M58530</td>
<td>A</td>
<td>Report with patient data viewed</td>
</tr>
</tbody>
</table>
Machine Learning for Burnout Prediction

**Goal:** A system that can assess the risk of burnout

Available in all hospitals & health centers

Electronic Health Records (EHR)

Activity Logs (digital footprint)

Feature Engineering

ML Model

Features
(e.g., EHR time, number of patients)

Burnout Outcome
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
Burnout Score

0.0  Low  4.0  High

EHR: Electronic Health Records

Association with Burnout

Statistical analysis shows: 
Higher workload links to higher risk of burnout.

e.g., 25% burned-out vs 10% healthy physicians have spent > 40H/month on "after-hour" working

There exists predictive information in EHR activity logs

Workload & Burnout
Study Design and Data Collection

- **Survey**
  - Sept / Oct / Nov
  - Measure burnout (PFI)
  - Month 1
  - Measure burnout (PFI)
  - Month 2
  - ... (Month 6)

- **EHR Audit Log**
  - Measure EHR use / mo
  - - Work hours
  - - Patient load
  - - Time spent on notes, chart review, inbox, etc.
  - - Wrong-patient orders
  - Measure EHR use / mo
  - Feb / Mar / Apr
Machine Learning Evaluation Pipeline

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

301 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

Training Set

Testing Set

Fold 1

Fold 2

Fold 3

Fold 4

Fold 5

Fold 6

Fold 7

Fold 8

Fold 9

Fold 10

Training Set (EHR data from 79-80 participants)

Testing Set (EHR data from 8-9 participants)

Normalization

Inner 10-fold CV Loop

Fold 1

Fold 2

Fold 3

Fold 4

Fold 5

Fold 6

Fold 7

Fold 8

Fold 9

Fold 10

Inner 10-fold CV used to identify optimal hyper parameters: cross-validated random search over parameter settings to maximize average AUROC over 10 folds

Best hyperparameters

10 evaluation score sets (AUROC & accuracy)

Train model using Training Set and evaluate on Testing Set

10x

Grouped Nested Cross-Validation

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</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>+ Workload</td>
<td></td>
<td></td>
<td></td>
<td></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*. 
Limitations

Study Design
• Small participant set (88 physicians x 6 months)
• Resident physicians may not be the right population to develop a stable model – not “steady state”
• Inter-individual variability in response to workload & workflow

Modeling
• Inability to capture non-EHR work, personal, environmental factors that also contribute to burnout
• Heavily require domain knowledge for feature engineering
Approach II: End-to-end Deep Learning
Directly Using Raw Logs

Proposal: End-to-end Burnout Prediction

**Goal:** A system that can *unobtrusively* monitor and assess the risk of burnout in real time

Available in all hospitals & health centers

---

Electronic Health Records (EHR)

Activity Logs
(digital footprint)

Deep Neural Networks

Running in Background (Unobtrusive)

Workload
(e.g., EHR time, number of patients)

Burnout Outcome
Data Collection

Model Training

Monthly surveys as burnout labels for model training

Data Collection

Survey

Activity Log

Shift 1 Shift 2

Month 1 Month 2 Month M

(till now, 88 intern physicians across 6 months)
Challenges & Proposed Model
Challenges to End-to-end Modeling

Challenge I
Content Understanding (Activity2Vec)

Challenge II
Large-scale Model Input

Challenge III
Limited Labels
I – Understanding Raw Activity Logs

Challenge I
Raw Content Understanding

Extract useful (data-driven) knowledge from raw activity logs.

- Over **1,900** unique kinds of actions
- **Irregular** intervals
- Need capture **dynamics/temporality**

No standard data encoding method for clinical activity logs.
Learning Activity Context

A: Note Review
B: Report Review

Semantically similar

Similar embedding (representation)

Embedding Lookup Table

~1,900 Vectors

Linear Projection

Neighbor Actions Target Action Neighbor Actions

“Skip-Gram Algorithm”
Learned Action Embedding

Identified Groups

Group 1: Mobile EHR client

Group 2: Preoperative assessment module

Group 3: Inpatient pre-rounding

Group 4: Outpatient setting
Learned Action Embedding

A. Most similar actions to 'Results Review accessed'

<table>
<thead>
<tr>
<th>Action</th>
<th>Similarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Results Review exited</td>
<td>0.779</td>
</tr>
<tr>
<td>Report viewed - Radiology Results</td>
<td>0.634</td>
</tr>
<tr>
<td>Report viewed - Result Detail</td>
<td>0.626</td>
</tr>
<tr>
<td>Report viewed - Lab Result Template</td>
<td>0.606</td>
</tr>
<tr>
<td>Report viewed - IP Vitals</td>
<td>0.582</td>
</tr>
<tr>
<td>Chart Review Encounters tab selected</td>
<td>0.557</td>
</tr>
<tr>
<td>Report viewed - CV Echo</td>
<td>0.551</td>
</tr>
<tr>
<td>Order list changed</td>
<td>0.550</td>
</tr>
<tr>
<td>Report viewed for an order</td>
<td>0.537</td>
</tr>
<tr>
<td>Report viewed - IP Flowsheet</td>
<td>0.529</td>
</tr>
</tbody>
</table>

B. Frequency distribution of average neighbor similarity
Action-time Joint Embedding

Contextual Action Embedding
\[ f_A(e_j) \]

Time Interval Embedding
\[ f_I(t_{i+1} - t_i) \]

Time Periodicity Embedding
\[ f_P(\sin(t_i)) \]

\[ e_j: \text{one-hot encoding of the j-th unique action} \]
\[ f: \text{Function implemented by a neural network} \]

\[ g_i = f(e_j, t_i) \]
Action-time Joint Embedding

\[ a_i^{(k)} = W_a e_i^{(k)} \]

\[ b_i^{(k)} = \tanh \left( W_b \cdot \log \left( t_i^{(k)} - t_{i-1}^{(k)} \right) + d_b \right) \]

\[ g_i^{(k)} = \text{Concat}([a_i^{(k)}; b_i^{(k)}; c_i^{(k)}]) \]

Action
Embedding

Time **Interval**
Embedding

Time **Periodicity**
Embedding

\[ c_i^{(k)}[j] = \begin{cases} 
\omega_j t_i^{(k)} + \varphi_j, & \text{if } j = 1 \\
\sin(\omega_j t_i^{(k)} + \varphi_j), & \text{if } 1 < j \leq d 
\end{cases} \]
II – Modeling on Long Dependencies

Challenge II
Large-scale Model Input

- 20,000 ~ 90,000 actions per month*person
- 1,000 ~ 8,000 actions per shift*person
- Need Long dependencies

*RNN: Recurrent Neural Networks

- "no-brainer" for sequential data
- Recurrent structure → slow to train
- Gradient issue → worse for long data sequence

Table: Statistics of Dataset

<table>
<thead>
<tr>
<th>Statistics</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td># total data points (actions)</td>
<td>15,767,634</td>
</tr>
<tr>
<td># types of actions</td>
<td>1,961</td>
</tr>
<tr>
<td># participants</td>
<td>88</td>
</tr>
<tr>
<td># total months of audit logs</td>
<td>754</td>
</tr>
<tr>
<td># months with eligible surveys (labels)</td>
<td>391</td>
</tr>
</tbody>
</table>

Model Training Time (per epoch)

<table>
<thead>
<tr>
<th>Model</th>
<th>Training Time (per epoch)</th>
</tr>
</thead>
<tbody>
<tr>
<td>RNNs</td>
<td>&gt; 5 h</td>
</tr>
</tbody>
</table>

RNN: Recurrent Neural Networks
Temporal clustering of activities may contain useful information associated with burnout.
Hierarchical Sequence Modeling

High-level Encoder:
Daily measurement $\rightarrow$ Monthly measurement

Low-level Encoder:
Activity embedding $\rightarrow$ Daily measurement

Raw data encoding:
Raw activity logs $\rightarrow$ Activity embedding (action & time)
Design Choice: Sequence Models

Recurrent Neural Networks (RNN) (e.g., LSTM, GRU)

- Compute steps recurrently
- Time-consuming
- Gradient issue for long sequences

Temporal Convolutional Networks (TCN)

- Compute steps concurrently
- Efficient for long sequences
- Behave like RNN, compute like CNN
Variations of TCN

(A) CausalNet

(B) ResTCN

(C) Fully Convolutional Networks (FCN)

\[ y = W \ast x + d \]
\[ z = \text{BatchNorm}(y) \]
\[ x' = \text{ReLU}(z) \]
III – Utilizing Unlabeled Data

Challenge III
Limited Labels

Action Pretraining
Embedding Lookup Table
Action (one-hot)
Neighboring Actions
Target Action
Neighboring Actions
Linear Projection

SeqEncoder Pretraining
SeqEncoder
Transfer
h
SeqDecoder
Transfer
e
Raw EHR Activity Logs

HiPAL
High-level Encoder
Low-level Encoder
LSTM
Tail Drop
Temporal Consistency Regularization
MLP
Monthly Measure

III
–
Utilizing Unlabeled Data

SeqEncoder

SeqEncoder Pretraining

SeqEncoder

SeqDecoder

Transfer

Monthly Measure

Shift 1

Action Pretraining

Time-dependent Activity Embedding

e^{(1)}
\cdots
e^{(k)}

Neighboring Actions
Target Action
Neighboring Actions
Experiment
**Prediction Performance**

**AUROC** (higher → better): Average true positive rate over all possible true negative rate

<table>
<thead>
<tr>
<th>Method</th>
<th>Training Time (per epoch)</th>
</tr>
</thead>
<tbody>
<tr>
<td>RNNs</td>
<td>&gt; 5 h</td>
</tr>
<tr>
<td><strong>HiPAL</strong></td>
<td>&lt; 1 min</td>
</tr>
</tbody>
</table>

Size of training set - 6 million activity logs
GPU - Nvidia RTX 3090
Practical Effectiveness

**AUROC = 0.6358**

% True Positive Rate  \( \rightarrow \)  % True Negative Rate

<table>
<thead>
<tr>
<th>Intervention</th>
<th>Cost</th>
<th>Model Preference</th>
<th>% True Positive</th>
<th>% True Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>Low</td>
<td>Sensitive</td>
<td>80%</td>
<td>41.5%</td>
</tr>
<tr>
<td>EHR training</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Online therapy</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>II</td>
<td>High</td>
<td>Specific</td>
<td>47.8%</td>
<td>80%</td>
</tr>
<tr>
<td>Days off</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vacations</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Training Efficiency

Single-level:
FCN, CausalNet and ResTCN

Hierarchical:
H-RNN: multi-level LSTM for video classification
HierGRU: replace $f_{Low}$ as GRU

Ours:
HiPAL-f: $f_{Low}$ = FCN
HiPAL-c: $f_{Low}$ = CausalNet
HiPAL-r: $f_{Low}$ = ResTCN
($f_{High}$ = LSTM)

RNN-based:
GRU and LSTM (recurrent, slow)
Temporal Consistency Regularization

Daily risk score:
\[ \alpha^{(k)} = \text{Softmax}(W_R \cdot r^{(k)} + d_R) \]

Daily loss:
\[ \mathcal{L}_L = \frac{1}{T} \sum_{k=1}^{T} \text{CE}(\alpha^{(k)}, y) \]

Monthly loss:
\[ \mathcal{L}_H = \text{CE}(\gamma, y) = y \log \gamma + (1 - y) \log(1 - \gamma) \]

Overall loss:
\[ \mathcal{L} = \mathcal{L}_H + \lambda \mathcal{L}_L \]
Can demonstrate how **daily risk** evolved over time and accumulated to final **monthly risk** of burnout.
Ablation Study

**TC:**
Temporal Consistency

**Pretrain:**
Action embedding pretraining (unsupervised)

**Concat → Add:**
Replace the embedding aggregation function from concatenation to numerical addition

<table>
<thead>
<tr>
<th>VARIANT</th>
<th>AUROC</th>
<th>AUPRC</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>w/o Regularization</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>w/ Different Embedding Settings</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>w/ All</td>
<td>$0.6347$</td>
<td>$0.5502$</td>
<td>$0.6400$</td>
</tr>
</tbody>
</table>
Ablation Study - Taildrop

![Graph showing AUC vs Prediction Time Offset (Day)]
Conclusion
Conclusion

• **End-to-end** deep learning-based burnout prediction framework
• Learn **contextual** representations directly from raw activities
• Capture natural **hierarchical** structure of clinician activity logs
• Model **long** dependencies on large-scale activity logs with **high efficiency**
• Facilitate fast burnout phenotyping and **individualized** intervention
Limitations

Study Design
• Small participant set (88 physicians x 6 months)
• Resident physicians may not be the right population to develop a stable model – not “steady state”
• Inter-individual variability in response to workload & workflow

Modeling
• Did not consider dependencies across months
• Lack of lower-level explainability
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
Related Published Work

Predicting Physician Burnout using Clinical Activity Logs: Model Performance and Lessons Learned
Sunny S. Lou, Hanyang Liu, Benjamin Warner, Derek Harford, Chenyang Lu, Thomas Kannappalli
Journal of Biomedical Informatics, 2022. [pdf]

Characterizing the Microstructure of EHR Work Using Raw Audit Logs: An Unsupervised Action Embeddings Approach
Sunny S Lou, Hanyang Liu, Derek Harford, Chenyang Lu, Thomas Kannappalli
Journal of the American Medical Informatics Association (JAMIA), 2022. [pdf]

HiPAL: A Deep Framework for Physician Burnout Prediction Using Activity Logs in Electronic Health Records
Hanyang Liu, Sunny S. Lou, Ben Warner, Derek Harford, Thomas Kannappalli, Chenyang Lu
ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), 2022. [pdf] [code] [presentation]
This study was funded by the Fullgraf Foundation, the Washington University/BJC HealthCare Big Ideas Innovation Award, and NIH 5T32GM108539-07.