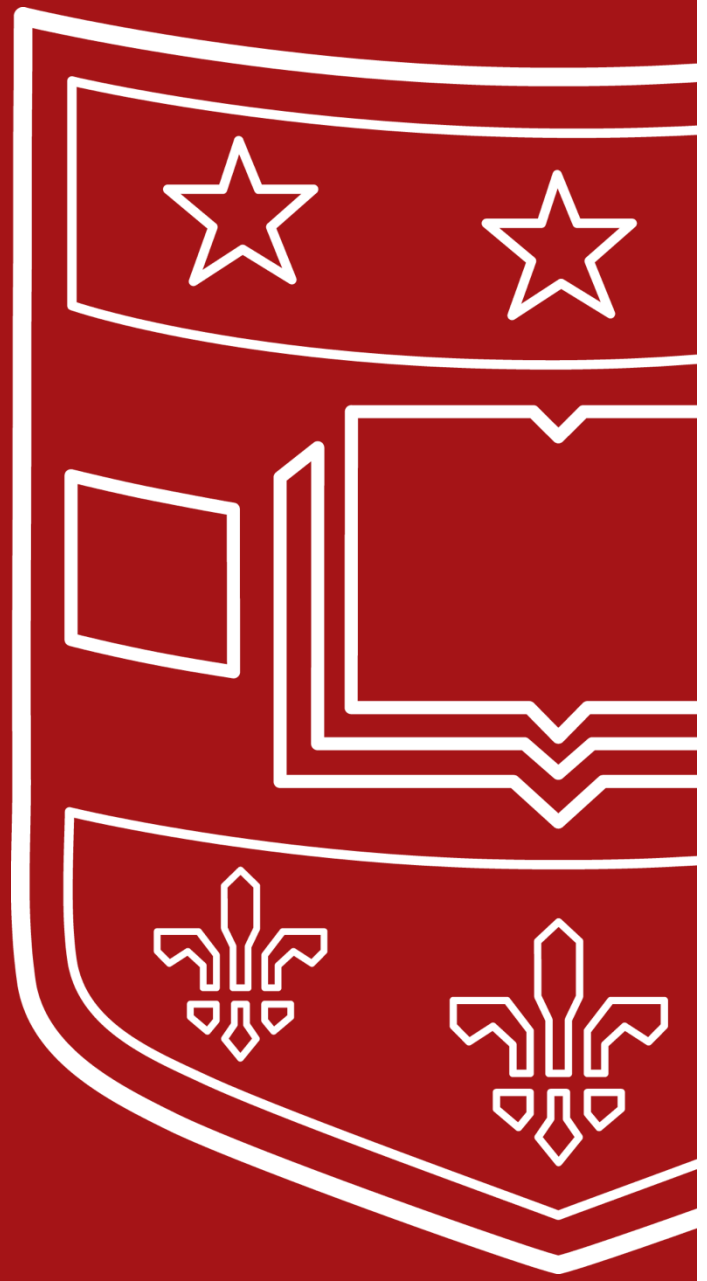
AI for Surgical Risk Prediction: Translating Data to Clinical Impact

Ziqi Xu
Ph.D. Candidate in Computer Science; AI for Health Institute
Advisor: Dr. Chenyang Lu
Sep 29th 2026



Washington University in St. Louis

The Missing Picture: Health Outside the Hospital

- Topics we covered so far
  - EHR
  - Clinical notes
  - Imaging
  - Genetics

- What happens outside the hospital?
  - Daily activity
  - Sleep
  - Heart rate
  - Mental states, mood, feeling

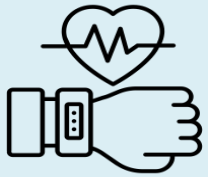


mHealth (Mobile Health)



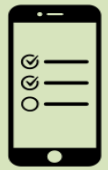
Collected in clinical settings ➡ **Life outside clinical visits**

Multimodal Data



Wearable Device

Fitbit
Step Count
Heart Rate
Sleep Time



Ecological Momentary Assessment (EMA)

Mobile Surveys
Self-report score of mental states



Electronic Health Records (EHR)

Tabular Records
Demographics
Medical History



Clinical Notes

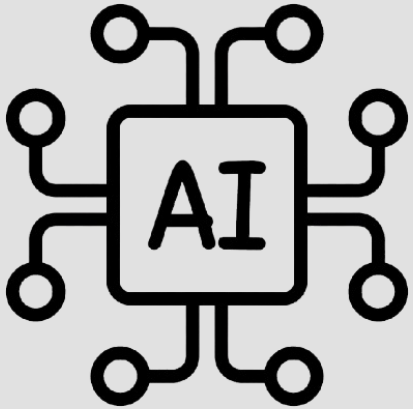
Free Text
Discharge Summary



Modalities are complementary, capturing different aspects of patients

Trustworthy AI

- AI has the transformative potential with the multimodal data
- However, clinical decision is high-risk in nature
- Many practical problem remains
  - Overconfident; Hallucination
- AI should be **transparent** about prediction and confidence/uncertainty



Human AI Collaboration



Challenges

- **Missing and noisy** data from real-world settings
  - Wearables: Fine-grained, noisy, incomplete time-series
- Integrating **heterogeneous** data sources
  - Data compliance and quality
- Building **trustworthy AI** for clinical use
  - Overconfidence in predictions and need to quantify uncertainty

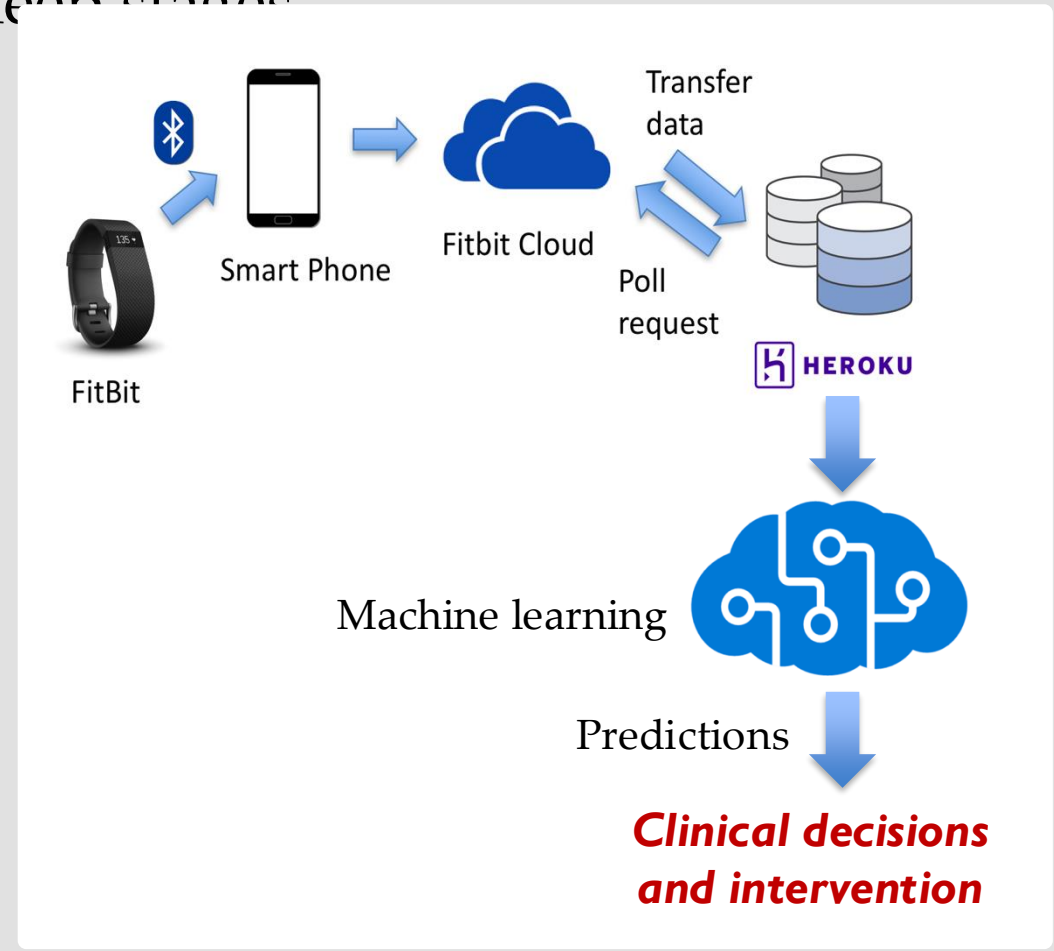
AI Applications in Surgical Risk Prediction

- **Pancreatic surgical outcome prediction with wearables**
  - **Continuous physiological signals**
- Multimodal Multitask lumbar recovery modeling
  - Clinical + behavioral + sensor data
- Uncertainty-aware post-surgical pain prediction
  - Confidence for clinical decisions
- Ongoing work and opportunities with GenAI, foundation models, etc.

Wearables

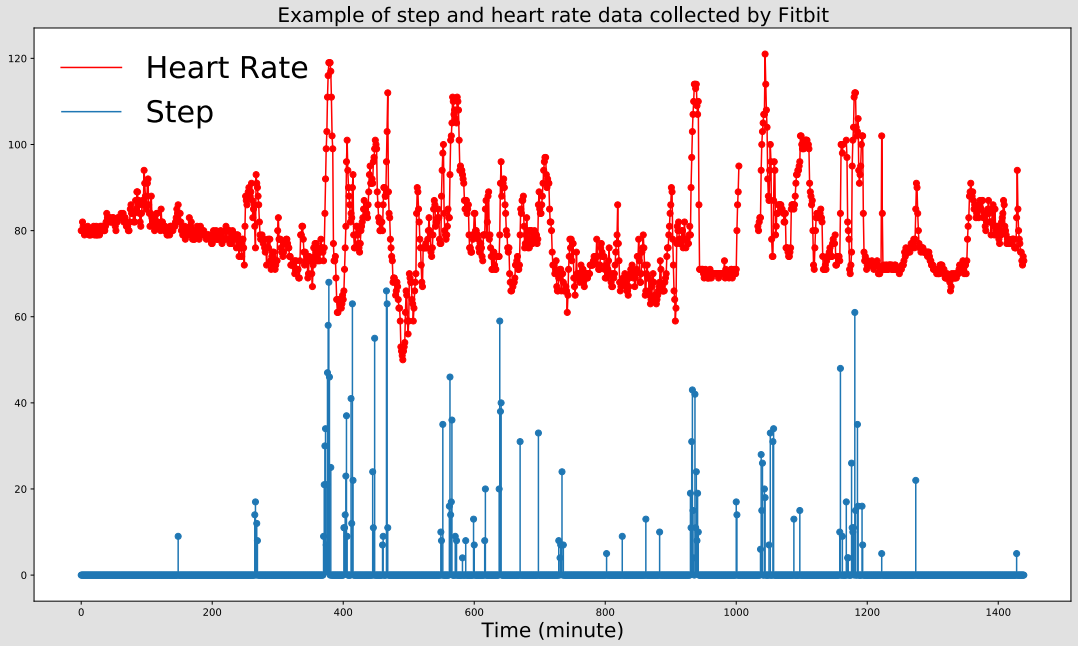
- Commonly available: step, heart rate, sleep stages
- More sensing modalities
  - Oxygen saturation (SpO2)
  - Skin temperature
  - Breathing rate
  - Heart rate variability
  - ECG
  - Stress
- **500+ million** wearables sold in 2021

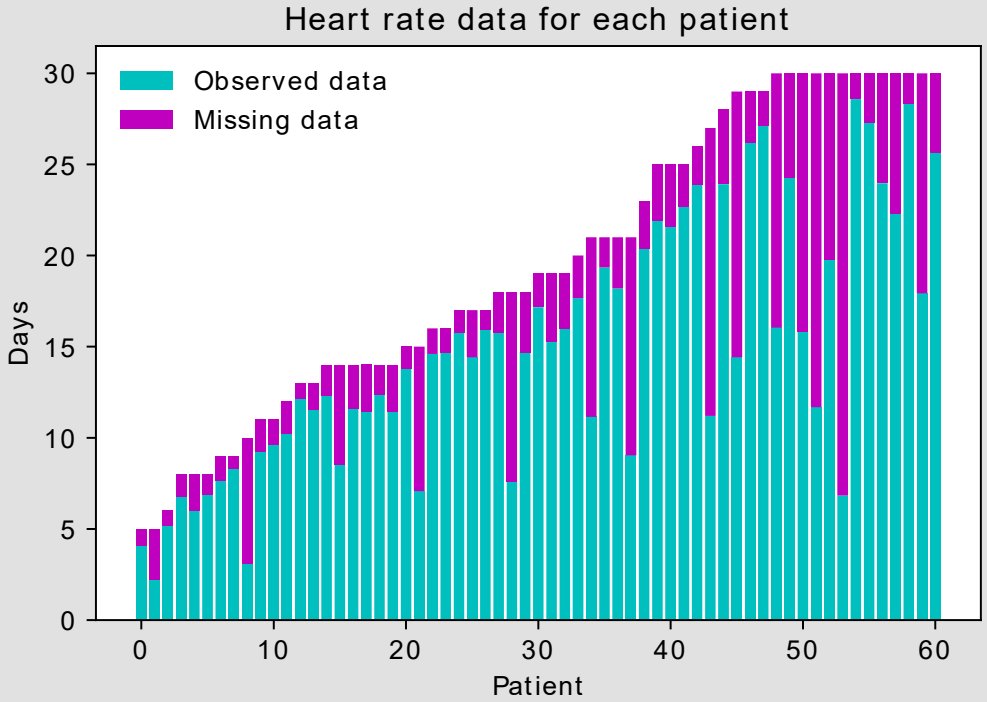
Unobtrusive, longitudinal monitoring outside hospitals!



Wearable Data

- Time series: step count, heart rate, sleep stage





***Fine-grained, noisy, incomplete* time-series**

Need: Predict Outcomes of Pancreatic Surgery

- Pancreatic cancer has a 5-year survival rate less than 5%.
- Surgery is a primary treatment but often followed by complications.
- Predict postoperative complications before surgery
  - **Decision support**: suitability for surgery
  - **Intervention**: prehabilitation

Preoperative Clinic **Surgery** **Outpatient Recovery**






Joint work with Chet Hammill (Surgery), Jingwen Zhang, Dingwen Li, Ruixuan Dai, Chenyang Lu (CSE)

Predict Postoperative Complications

- Goal: predict postoperative complications with preoperative data
  - Wearable data collected by Fitbit wristband
  - Patient clinical characteristics from medical record

- Prospective study: 61 patients undergoing pancreatic surgery
  - 25 (41%) experienced complications

- Machine learning approach
  - Small cohort → avoid deep models
  - Features extracted from wearable data → shallow machine learning models
  - Wearable data + Clinical characteristics

Robust Feature Engineering Pipeline



Daily Feature Extraction

Step Count

Heart Rate

Sleep Stage

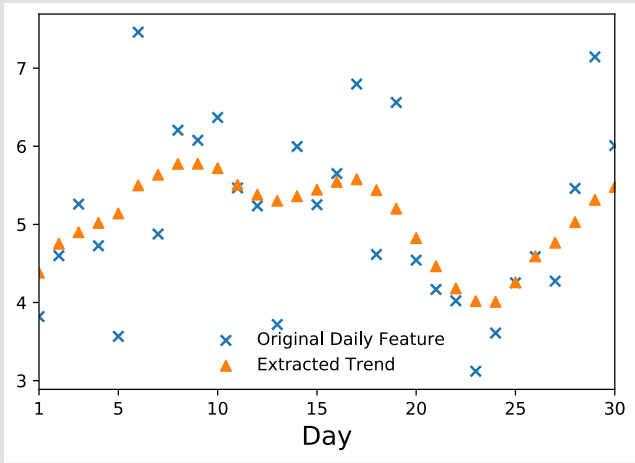
Wearable Time Series

N

D

D: Number of Daily Features
N: Number of Days

High-level Feature Extraction





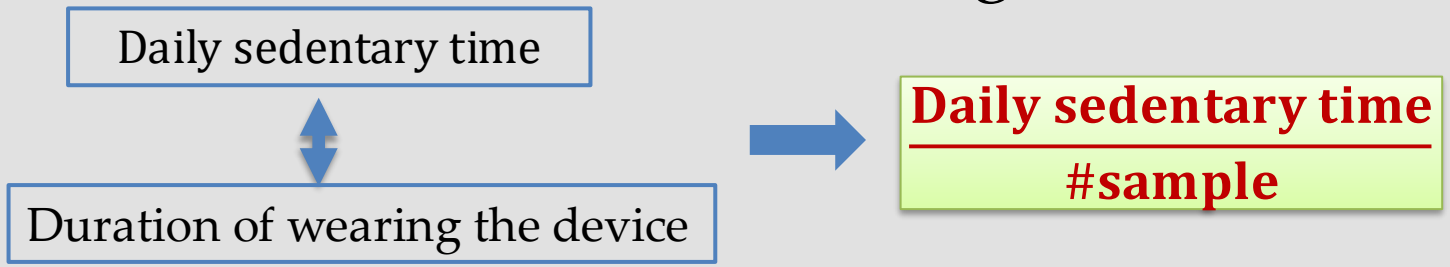
Clinical Characteristics

\+

Inputs to predictive model

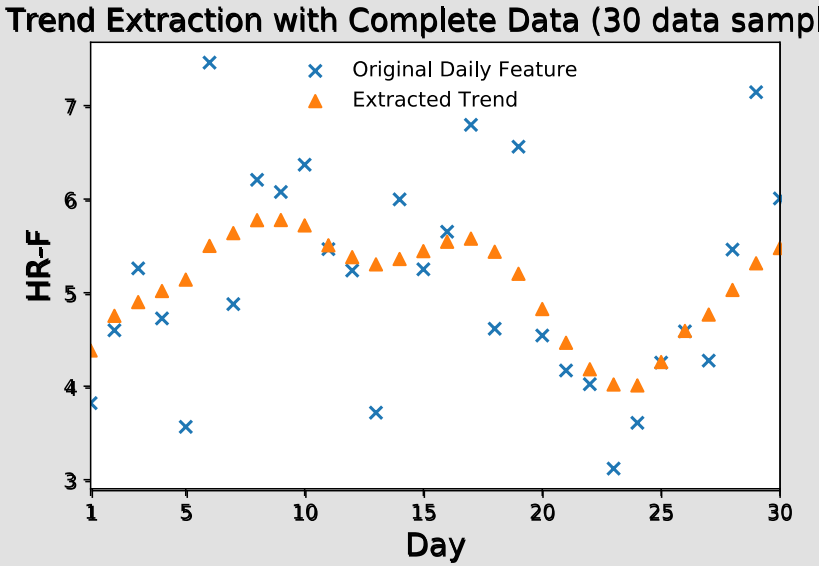
Raw Time Series → Daily Features

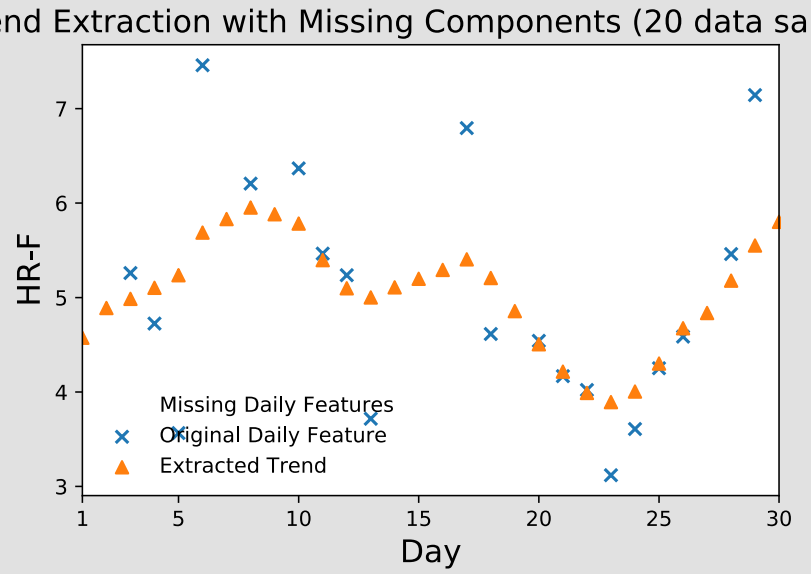
- **Statistical** features
  - max, sum, skewness, kurtosis, energy, entropy, inertia…
- **Semantic** features
  - Activity: daily steps, sedentary time, sedentary bout counts…
  - Sleep: awake counts, sleep efficiency, time of sleep, time before falling asleep…
- **Standardization**: account for missing data



Daily Features → High-level Features

- Singular Spectrum Analysis (SSA)
  - **denoise** time series of daily features.
  - **impute** missing daily features.





- **mean**
- **standard deviation**
- **slope**

Robust Prediction of Surgical Complications

- **Machine learning models outperform standard surgical risk scores.**
  - **x2** AUPRC
  - **x3** sensitivity at the same specificity

Data Source	AUROC	AUPRC	Sensitivity	Specificity
Random weighted classifier	0.5097 (0.0585)	0.4322 (0.0469)	0.1520 (0.0854)	0.8583 (0.0504)
NSQIP with Clinical Characteristics	0.6114 (0.0000)	0.4075 (0.0000)	0.2800 (0.0000)	0.8571 (0.0000)
ML with Clinical Characteristics	0.7632 (0.0085)	0.7374 (0.0206)	0.5800 (0.0699)	0.8583 (0.0083)
Wearable Data	0.7326 (0.0074)	0.7192 (0.0154)	0.5480 (0.0440)	0.8583 (0.0083)
Clinical Characteristics + Wearable Data	**0.8802 (0.0050)**	**0.8871 (0.0087)**	**0.8320 (0.0160)**	0.8583 (0.0083)

- NSQIP: American College of Surgeons National Surgical Quality Improvement Program
- AUROC: Area Under the Receiver Operating Characteristic Curve
- AUPRC: Area Under the Precision-Recall Curve

Robust Prediction of Surgical Complications

- Machine learning models outperform traditional surgical risk scores.
- **Wearables & clinical characteristics provide complementary information.**

Data Source	AUROC	AUPRC	Sensitivity	Specificity
Random weighted classifier	0.5097 (0.0585)	0.4322 (0.0469)	0.1520 (0.0854)	0.8583 (0.0504)
NSQIP with Clinical Characteristics	0.6114 (0.0000)	0.4075 (0.0000)	0.2800 (0.0000)	0.8571 (0.0000)
ML with Clinical Characteristics	0.7632 (0.0085)	0.7374 (0.0206)	0.5800 (0.0699)	0.8583 (0.0083)
Wearable Data	0.7326 (0.0074)	0.7192 (0.0154)	0.5480 (0.0440)	0.8583 (0.0083)
Clinical Characteristics + Wearable Data	**0.8802 (0.0050)**	**0.8871 (0.0087)**	**0.8320 (0.0160)**	0.8583 (0.0083)

- NSQIP: American College of Surgeons National Surgical Quality Improvement Program
- AUROC: Area Under the Receiver Operating Characteristic Curve
- AUPRC: Area Under the Precision-Recall Curve

Summary: Predict Postoperative Complications



- Extract **high-level** features from **fine-grained** wearable data
  - Raw time series (minute) → Daily features
  - Time series of daily features → High-level features

- Extract **robust** features from **noisy** and **incomplete** wearable data
  - Standardized daily features from incomplete time series
  - Singular spectrum analysis for high-level feature extraction

J. Zhang, D. Li, R. Dai, H. Cos, G.A. Williams, L. Raper, C.W. Hammill, and C. Lu, Predicting Post-Operative Complications with Wearables: A Case Study with Patients Undergoing Pancreatic Surgery, Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (UbiComp'22), 6(2), Article 87, July 2022. **Distinguished Paper Award**

AI Applications in Surgical Risk Prediction

- Pancreatic surgical outcome prediction with wearables
  - Continuous physiological signals
- **Multimodal Multitask lumbar recovery modeling**
  - **Clinical + behavioral + sensor data**
- Uncertainty-aware post-surgical pain prediction
  - Confidence for clinical decisions

Predicting Lumber Spine Surgical Outcomes



- Burden of Low Back Chronic Pain
  - Affects life of 1.36 million people and cost $135 billion in USA
  - Lumbar spine surgery is potential treatment for that

- **Heterogeneous** and **Multidimensional** surgical recovery
  - People might be worse after taking the surgery
  - Interrelated aspects: Pain, Physical Function, Mental Health

Goal: Use **multimodal** pre-operative data to predict the **multidimensional** post-surgical recovery

Multi-modal Features



Correlation between EMA and Fitbit

Activity Induced Pain





Pain

Catastrophizing

Disability

Depression

Dynamic Subjective



Smartphone EMA Survey





Dynamic Objective



Static

Traditional Clinical Data

Fitbit

Activity



Heart Rate



Sleep



Multidimensional Surgery Outcomes

- **Heterogeneous** and **Multidimensional** surgical outcome



Δ Pinter Post- minus pre-surgical Pain Interference Score



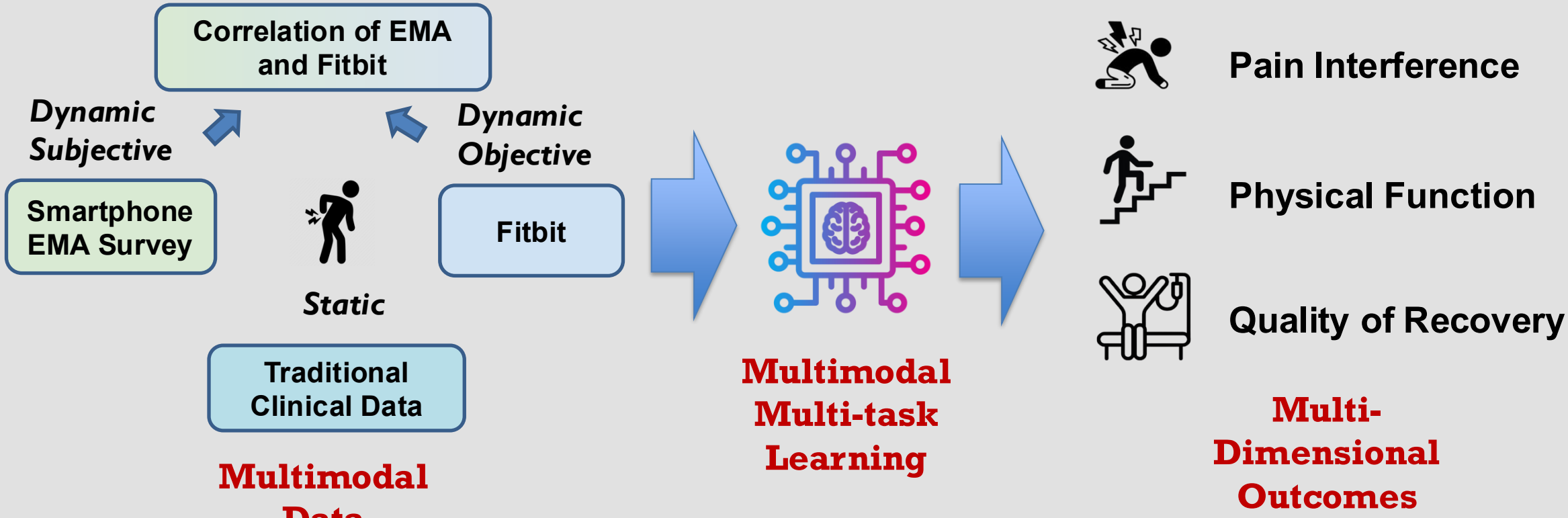
Δ PhysFun Post- minus pre-surgical Physical Function Score



QOR Quality of Recovery (General well-being)

Multitask learning exploits the **inherent correlations** among outcomes to potentially improve overall performance.

Predicting Lumbar Spine Surgical Outcomes



Xu Z, Zhang J, Greenberg J, Frumkin M, Javeed S, Zhang J, Benedict B, Botterbush K, Rodebaugh T, Ray W, Lu C. Predicting multi-dimensional surgical outcomes with multi-modal mobile sensing: A case study with patients undergoing lumbar spine surgery. *Proc ACM Interact Mob Wearable Ubiquitous Technol.* 2024;8(2).

Multimodal Multitask Feature Selection



- **Challenge I: Feature Selection**
  - From each different modalities to have comprehensive representation over patients
    - Avoid dominance of single modality
  - Informative to all the tasks

- **Using Multitask Group Lasso Feature Selection [2]**
  - Key idea: Incorporate multitask loss and with grouping of features
  - Treat each modality as a feature group
  - Within each group:
    - identify features that are informative across tasks
  - The group structure naturally accounts for modality dimensionality
  - Multitask loss: select features that contribute jointly to predicting all outcomes

[2] Ming Yuan and Yi Lin. 2006. Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society. Series B: Statistical Methodology 68, 1 (2006), 49–67

Multitask Learning with Dynamic Weights



- Motivation: Leverage labels from several outcomes together to improve performance
  - Surgical recovery with pain, physical function and quality of recovery are interrelated

$$L_{reg} = \frac{1}{N}\sum_{t=1}^{T}\sum_{n=1}^{N} w_t \cdot MSE(y_{n,t}, \hat{y}_{n,t})$$

w_t: the weights for each task

If NOT properly set,
it can introduce **negative transfer**

- Using **uncertainty** to automatically weigh loss [3]
  - Key idea: Large uncertainty for one task will lower its contribution.
  - Model the task regression prediction as a Gaussian distribution:

$$p(\mathbf{y}_1, \mathbf{y}_2 | \mathbf{f}^{\mathbf{W}}(\mathbf{x})) = \mathcal{N}(\mathbf{y}_1; \mathbf{f}^{\mathbf{W}}(\mathbf{x}), \sigma_1^2) \cdot \mathcal{N}(\mathbf{y}_2; \mathbf{f}^{\mathbf{W}}(\mathbf{x}), \sigma_2^2).$$

$$= \frac{1}{2\sigma_1^2}\mathcal{L}_1(\mathbf{W}) + \frac{1}{2\sigma_2^2}\mathcal{L}_2(\mathbf{W}) + \log \sigma_1 \sigma_2$$

Tune w_t into a trainable parameter: σ

[3] A. Kendall, Y. Gal, R. Cipolla, Multi-Task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics, Cvpr. (2018) 7482–7491.

Case Study

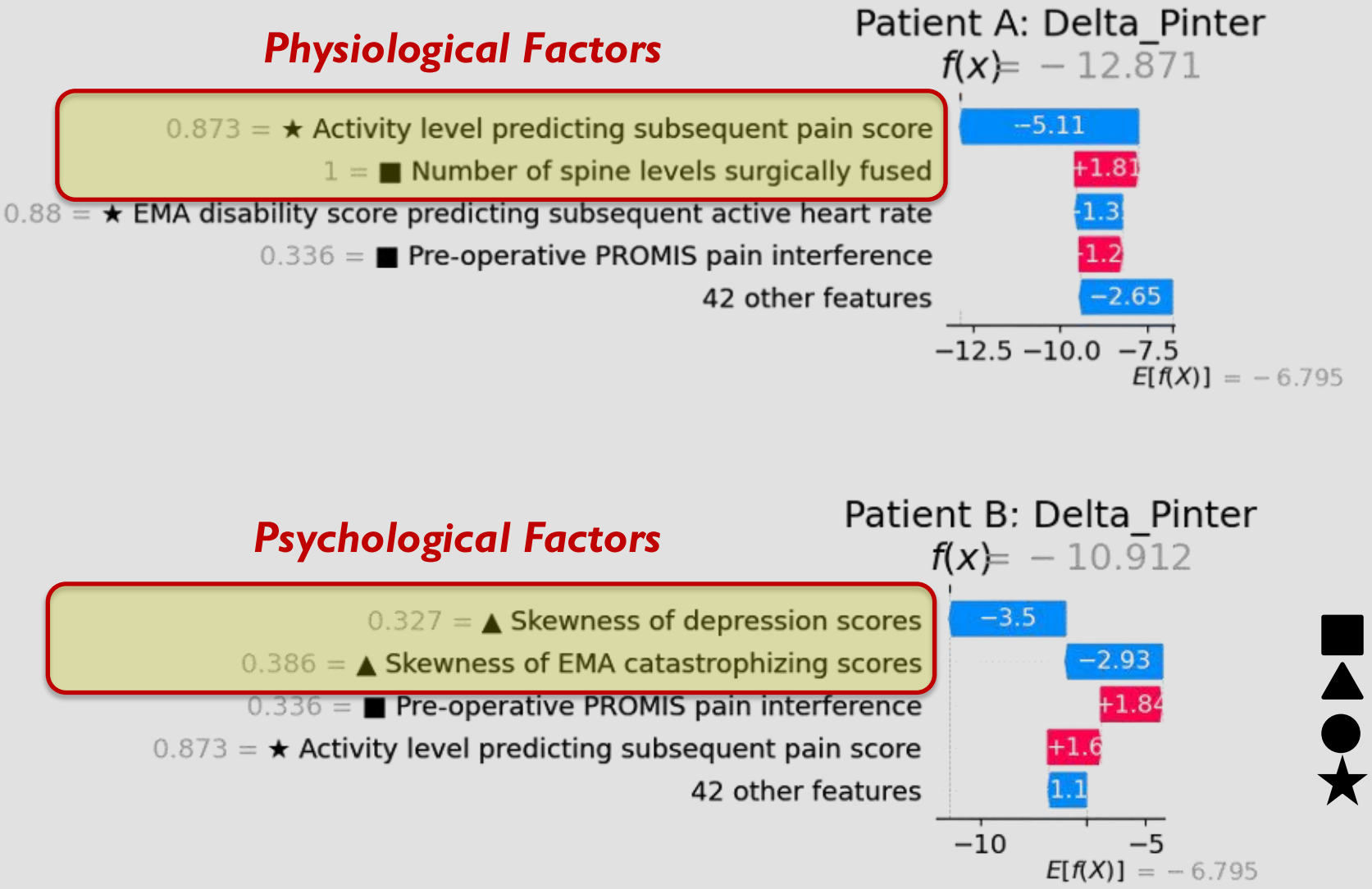
- 122 lumbar spine surgery conducted at a large academic medical center
- Answer EMAs using smartphones
- Wear Fitbit as much as possible

	Δ Pinter (Range 43.4)		Δ PhysFun (Range 38.1)		QOR (Range 102)	
	RMSE	Adjusted R^2	RMSE	Adjusted R^2	RMSE	Adjusted R^2
Best Single-Task	7.6903	0.1785	6.6969	0.0573	18.1399	0.0707
Traditional Multi-Task	8.3126	0.0401	6.6828	0.0613	17.4709	0.138
Our Approach	**7.4462**	**0.2298**	**6.4245**	**0.1212**	**14.9226**	**0.3711**

- **RMSE**: Root Mean Square Error
- **Adjusted R^2**: R-square accounting for number of predictors

- **STL** fails to predict complex recovery outcomes.
- **Traditional MTL** leads to negative transfer among tasks.
- Our Approach achieves the best performance overall.

Personalized Model Explanation



Multimodal Contribution

Personalized Treatments

■ : Traditional Clinical Data
▲ : Smartphone EMA Survey
● : Fitbit Monitoring
★ : Correlation of EMA and Fitbit

Summary: Predict Post-surgical recovery

- Multimodal Multitask Learning
  - Integrate clinical, survey, and wearable data
  - Capturing complementary signals across data sources
  - Jointly predict multiple recovery outcomes with multitask learning

- Personalized prediction and treatments
  - Identify key factors driving individual outcomes
  - Quantify contributions from different data sources
  - Enable more tailored clinical decision-making

Xu Z, Zhang J, Greenberg J, Frumkin M, Javeed S, Zhang J, Benedict B, Botterbush K, Rodebaugh T, Ray W, Lu C. Predicting multi-dimensional surgical outcomes with multi-modal mobile sensing: A case study with patients undergoing lumbar spine surgery. Proc ACM Interact Mob Wearable Ubiquitous Technol. 2024;8(2).

AI Applications in Surgical Risk Prediction

- Pancreatic surgical outcome prediction with wearables
  - Continuous physiological signals
- Multimodal Multitask lumbar recovery modeling
  - Clinical + behavioral + sensor data
- **Uncertainty-aware post-surgical pain prediction**
  - **Confidence for clinical decisions**

Predicting Postsurgical Persistent Pain (PPSP)

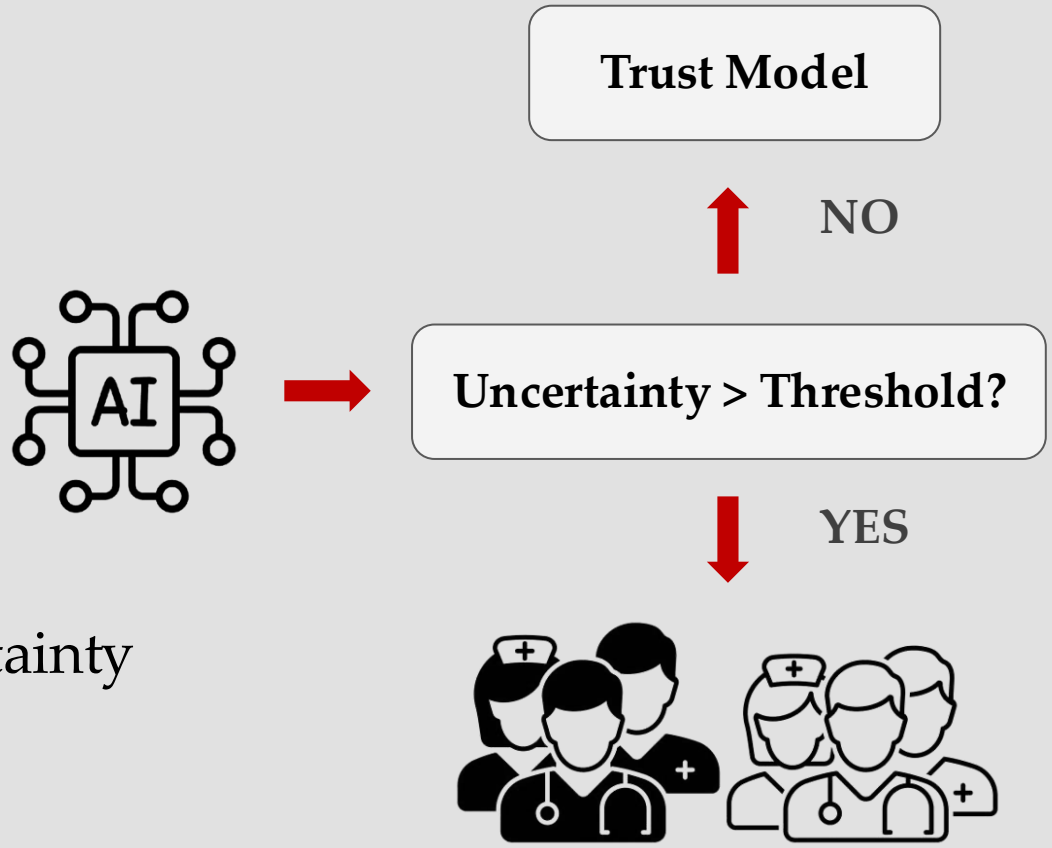


- Persistent postoperative pain (PPSP)
  - Pain that persists for more than 3 months after surgery
  - One of most common complications
  - Affect 45 million people each year and significant financial burden
- Prior studies use clinical data in hospital with simple machine learning
  - Lack assessments of patients' dynamic mental states outside hospital
  - Fails to assess uncertainty behind machine learning decision

Goal: Use **multi-modal** pre-operative data to
to predict PPSP with **uncertainty-aware** model

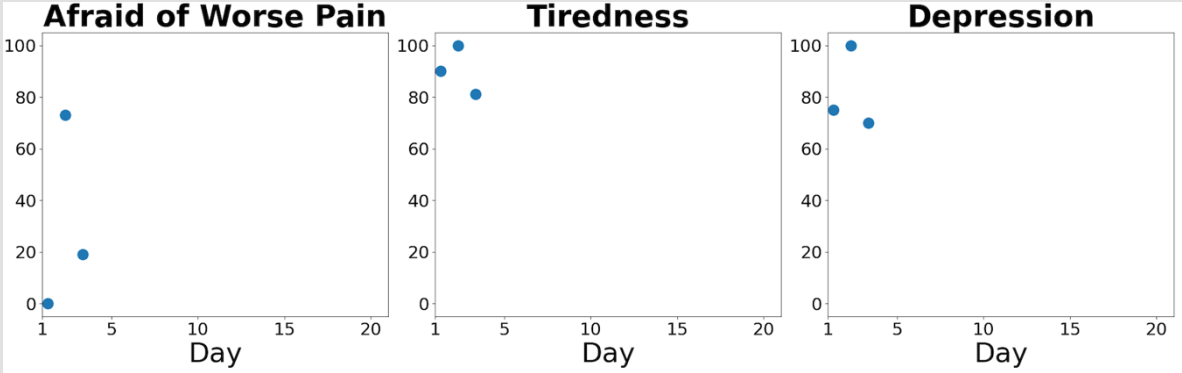
Why is uncertainty so important?

- Healthcare: high-risk in nature
  - Overconfident; Hallucination



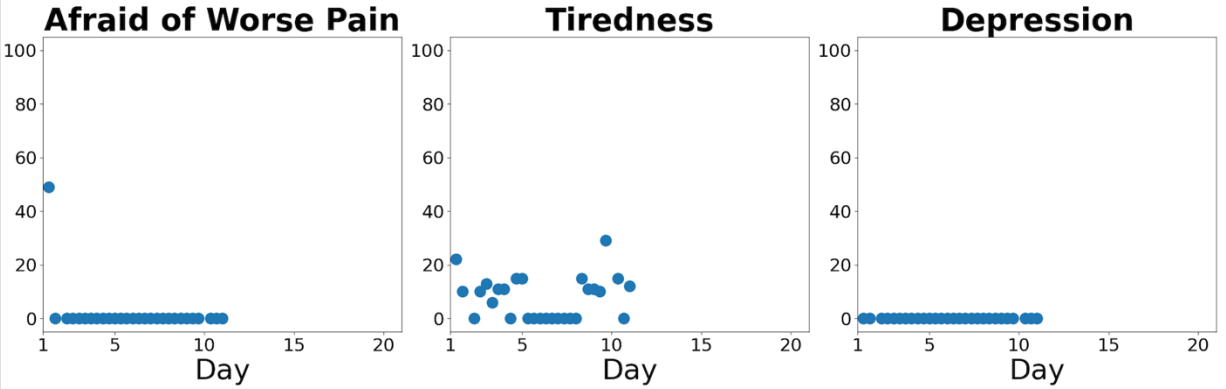
- **Trustworthy** AI:
  - Transparent about prediction and uncertainty
  - Communicate uncertainty well

Uncertainty with EMA as example



Patient A:
EMA uncertainty = 0.9939

Lower Compliance -> Less Information -> Higher Uncertainty



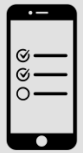
Patient B:
EMA uncertainty = 0.6892

Higher Compliance -> More Information -> Lower Uncertainty

Uncertainty-aware Multimodal Predictive Framework



EMA



Laboratory Tests



Vital Signs



Diagnoses



Demographics & Study Records



Evidential Deep Learning

Single-modality Prediction

Single-modality Uncertainty

Single-modality Prediction

Single-modality Uncertainty

Dempster–Shafer Theory for fusion

Combined Prediction

Combined Uncertainty

Lower Uncertainty -> Higher contribution
Higher Uncertainty -> Lower contribution

Evidential Deep Learning (EDL)



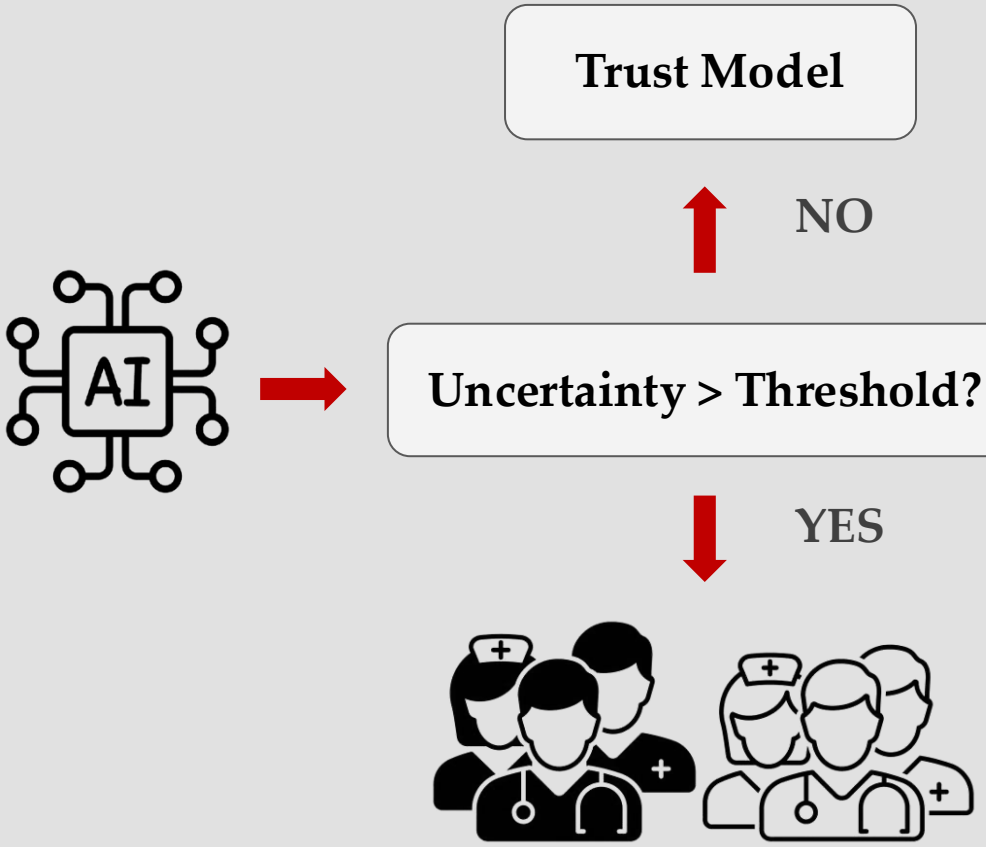
- Belief masses b_k ($k \in [1,\ldots K]$) for each class [4]
  - Higher values indicate a higher likelihood of belonging to one class
- Uncertainty mass u to indicate how much model does not know
  - Higher values indicate greater uncertainty

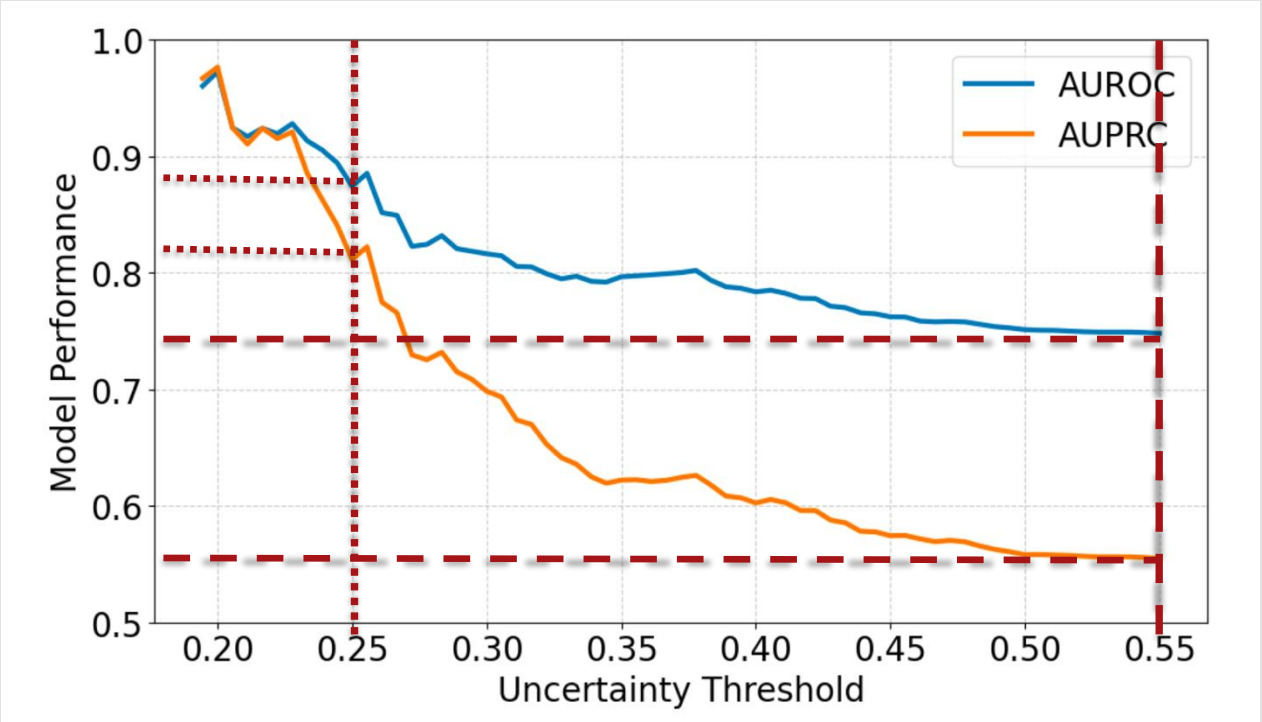
$$1 = u + \sum_{k=1}^{K} \mathrm{b_k}$$

Learnable from Neural Networks

[4] Sensoy, Murat, Lance Kaplan, and Melih Kandemir. "Evidential deep learning to quantify classification uncertainty." Advances in neural information processing systems 31 (2018).

Uncertainty measurements evaluation





Model performance significantly **increase** while **excluding patients with higher uncertainty.**

Summary: Predict Post-surgical recovery

- Uncertainty-Aware Multimodal Prediction
  - Integrate multiple data sources with varying reliability
  - Quantify model confidence alongside predictions
  - Adapt model decisions based on data quality
- Reliable and Actionable Predictions
  - Identify patients with uncertain predictions
  - Improve decision-making by incorporating uncertainty
  - Enable more trustworthy clinical use of AI

Xu Z, Zhang J, Haroutounian S, Liu H, Cao Z, Messner GR, Alaverdyan HB, Ahuja S, Koshy R, Hanns J, Frumkin M, Rodebaugh TL, Lu C. Incorporating uncertainty in predictive models using mobile sensing and clinical data: a case study on persistent post-surgical pain. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*. 2025;9(2):58.

Ongoing Work: Surgical Prehabilitation and Readiness (SPAR)

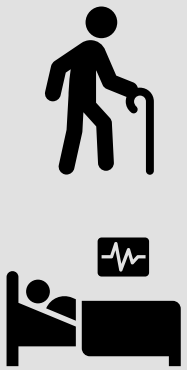
- **Heterogeneous** Patients & Surgeries
  - Varying procedures and patient subgroups
  - Need for robust, generalizable models

- From Controlled Trials to **Standard Care**
  - Wearable compliance in real-world settings
  - Highly Missing and noisy wearable data

- From Prediction to **Intervention**
  - Move beyond machine learning predictions to actionable insights
  - Enable targeted, personalized clinical treatments

Surgical Prehabilitation and Readiness (SPAR)

WashU/BJC's preoperative optimization program



Older (>70 years), high-risk
Inpatient surgery
≥2 weeks of enrollment

Physical activity:
- Physical therapy and telemedicine
- Remote activity tele-monitoring

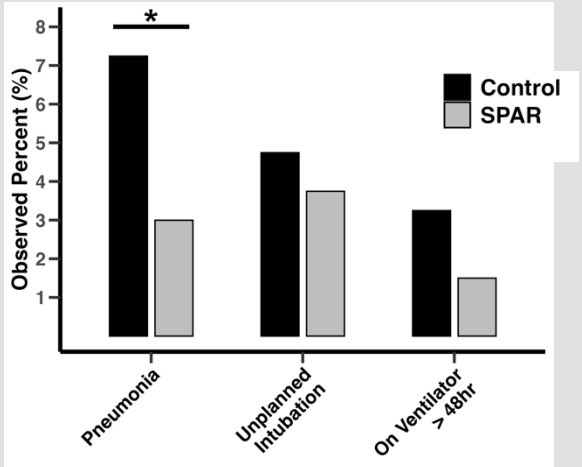
Respiratory fitness:
- Incentive spirometry
- Smoking cessation

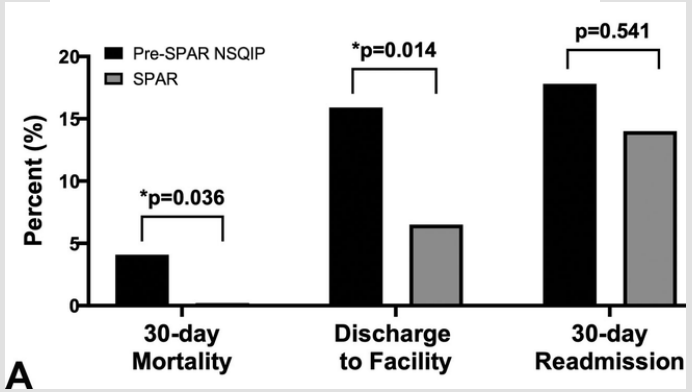
Nutritional optimization:
- Nutritional evaluation with registered dietician

Mindfulness:
- Gratitude journaling, meditation, sleep hygiene

4 domains targeted

SPAR patients do better than matched controls





Reduced postoperative pneumonia

Reduced 30-day mortality

Reduced discharge to facility

Data from Zarate Rodriguez et al., *J Am Coll Surg*, Oct 2023

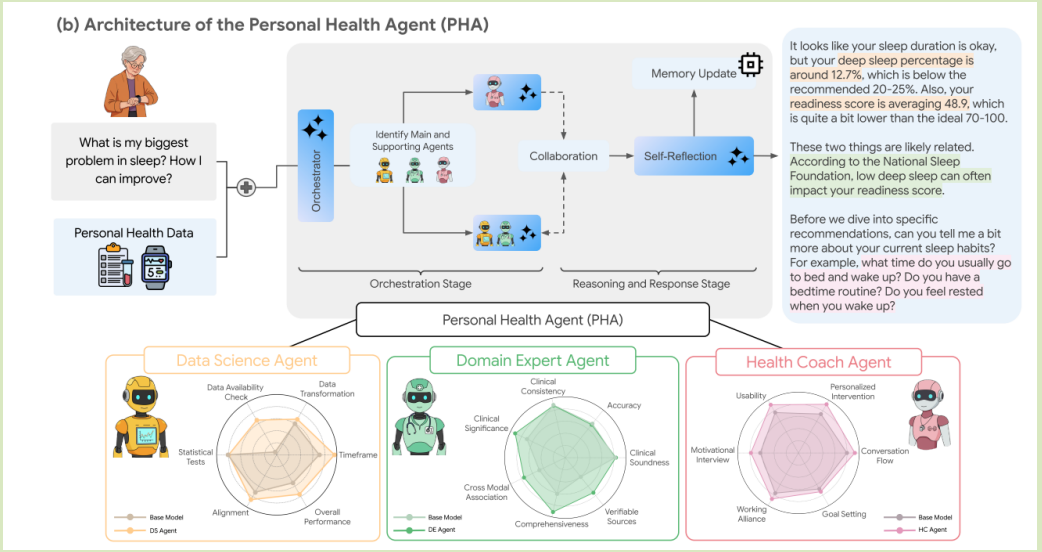
Future Opportunities: From Predictive AI to Generative AI



Wearable / Time-Series Foundation Models

- Wearables:
  - Google: SensorLM
  - Apple: Wearable Foundation Models
- General Time Series Foundation Models
  - Amazon: Chronos-2
  - Google: TimesFM
  - CMU: MOMENT

Agentic AI + Wearables [*]



* Heydari, A. Ali, et al. "The anatomy of a personal health agent." *arXiv preprint arXiv:2508.20148* (2025).

Conclusion and Takeaways

- AI Enables Richer Risk Prediction
  - Integrate multimodal data beyond traditional clinical variables
  - Capture complex, real-world patient signals
- Trustworthy Multimodal AI
  - Quantify uncertainty to reflect model confidence
  - Provide reliable and interpretable predictions for clinical use
- Clinical Value of AI
  - Improve risk stratification and decision-making
  - Enable more personalized patient care