AI for Surgery

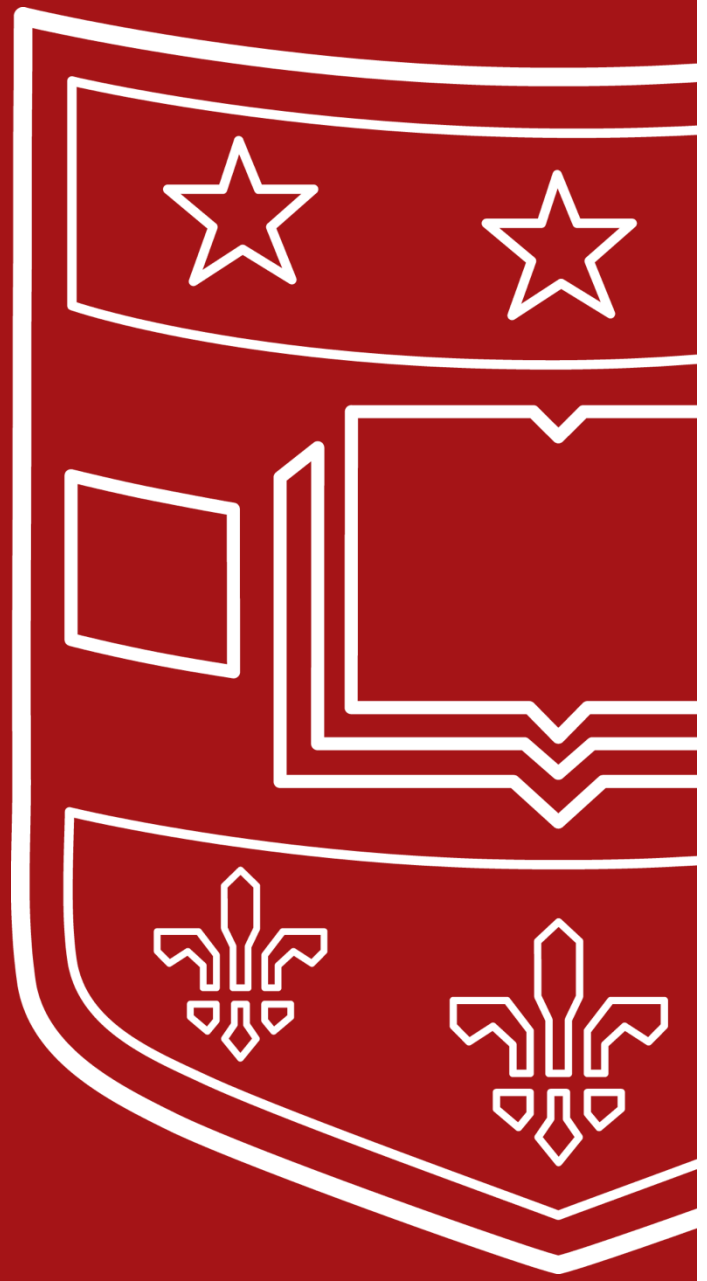
From Multimodal Data to Clinical Decisions

Chenyang Lu

Director, AI for Health Institute (AIHealth)

Fullgraf Professor of Computer Science,
Anesthesiology, Medicine, Neurosurgery, Public Health

Washington University in St. Louis



AI for Surgery

PHENOTYPE
complex diseases

DETECT EARLIER
identify risk sooner

PREDICT OUTCOMES
improve prognostication

GUIDE DECISIONS
support personalized care

CLINICAL DATA

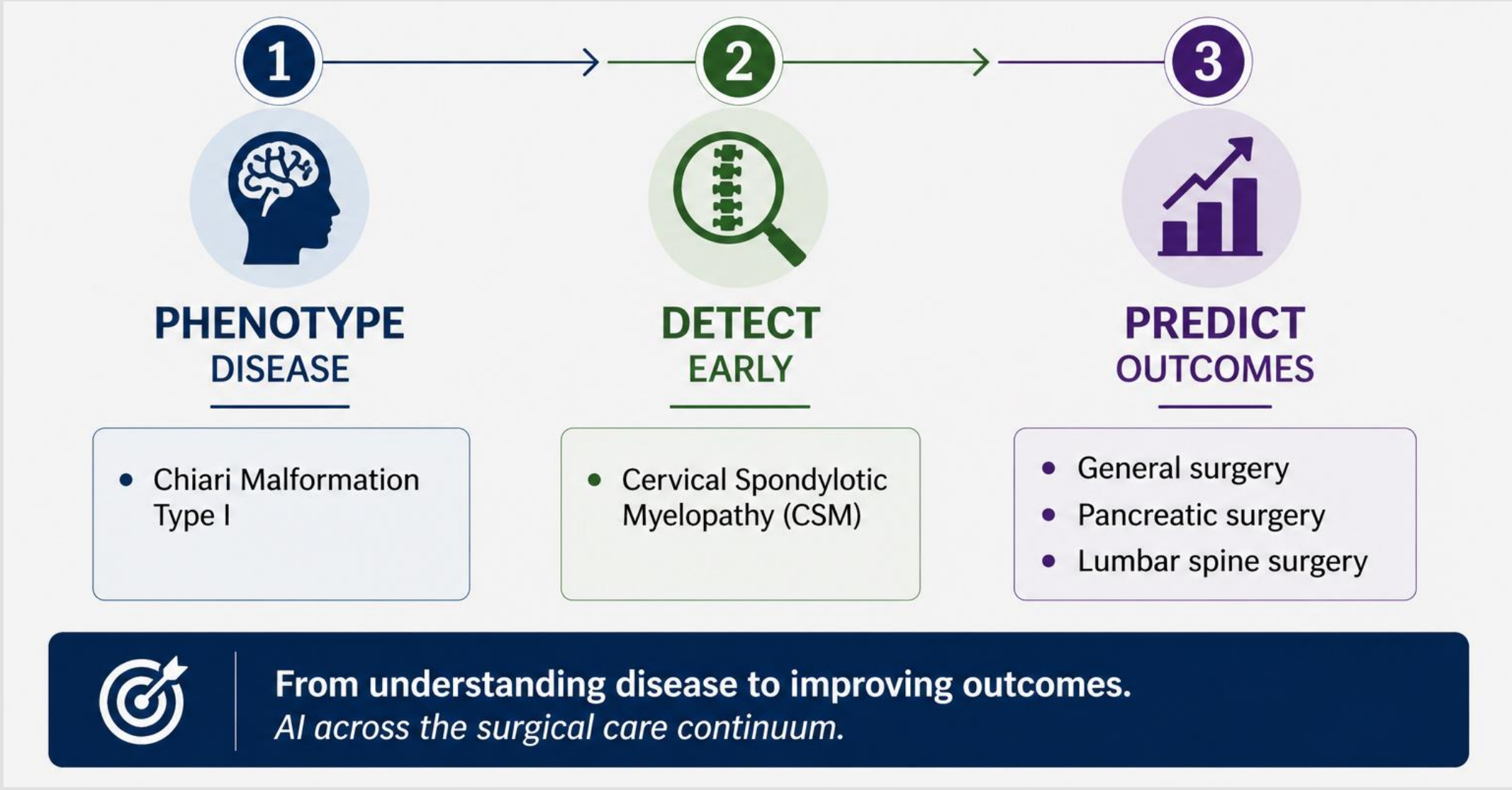
- EHR
- Clinical notes
- Imaging
- Genomics

PATIENT-GENERATED DATA

- Wearables
- Smartphones

AI transforms data into actionable insights to improve surgical care and patient outcomes.

AI for Surgery: Roadmap

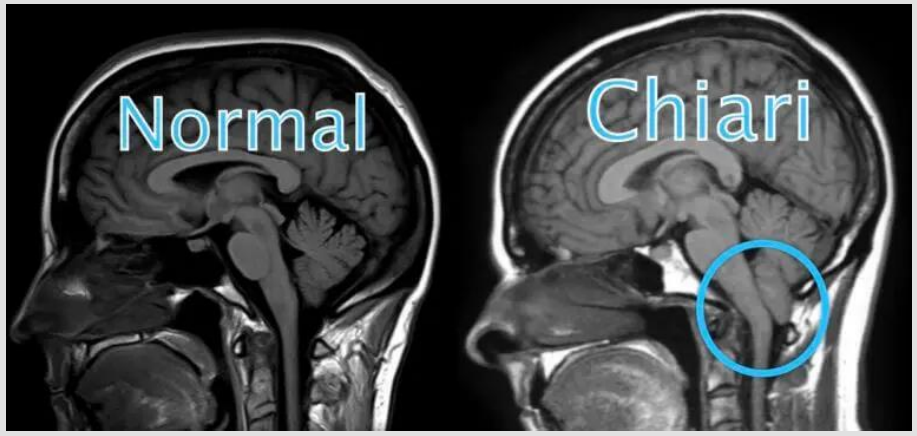


Chiari Type-1 Malformation (CM1)

- CM1 is a **heterogeneous** disease
  - an abnormality in the back of the head where the brain and spinal cord meet
  - ≥5 mm of cerebellar tonsillar ectopia below the foramen magnum on brain MRI scan

- Congenital disease affecting 3.6% of the population

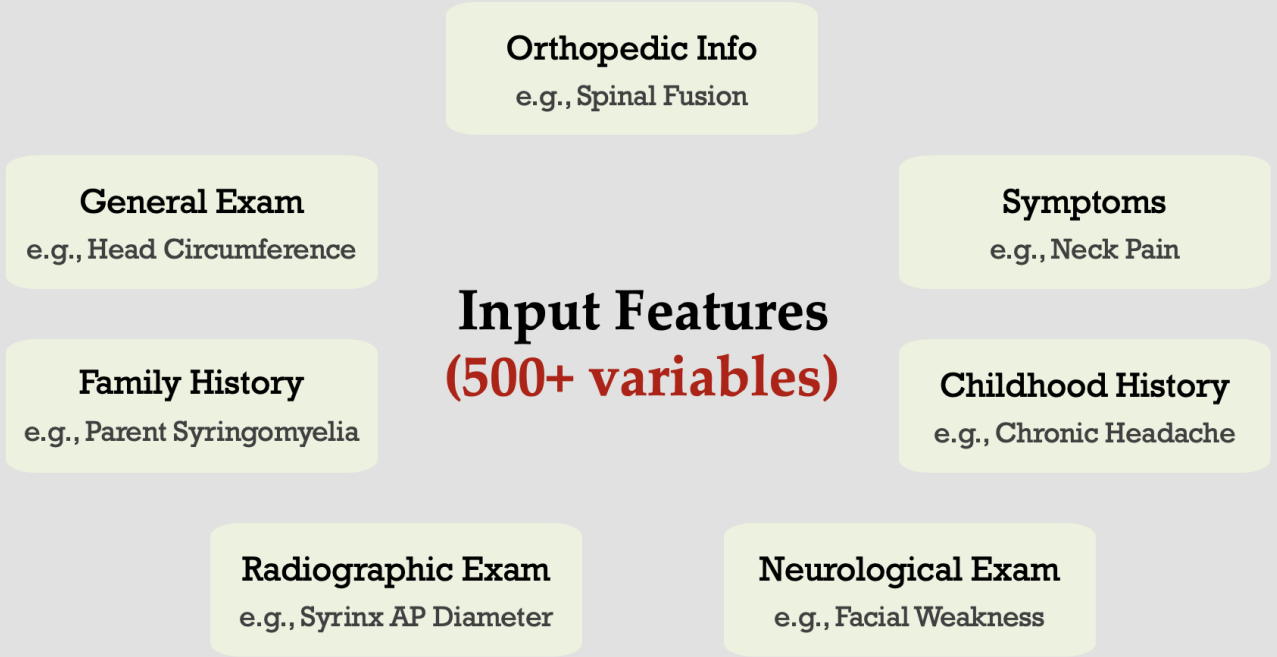
- **Significant differentiation in clinical presentation and natural history**



Phenotype Chiari Malformation Type 1 (CM1)

- CM1 likely reflects multiple clinical phenotypes rather than a single disease entity.
- Goal: Identify **distinct, clinically meaningful** phenotypes

Park Reeves Syringomyelia Research Consortium registry: 1340 patients from 42 medical centers since 2011



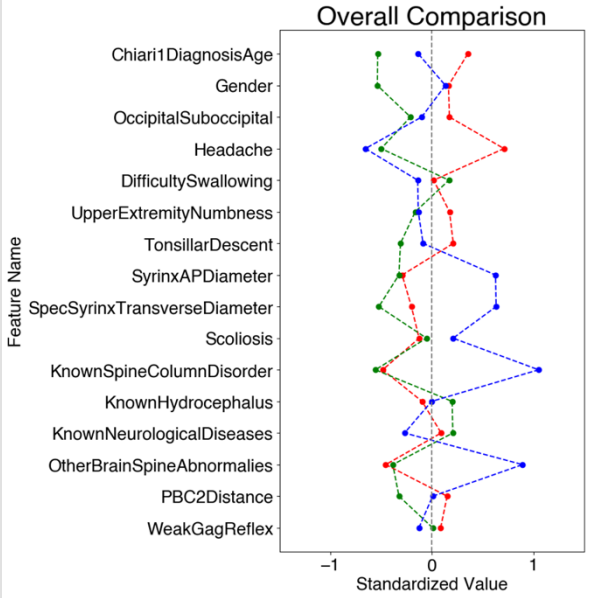
Work with Dave Limbrick, Jacob Greenberg, Sean Gupta (Neurosurgery), Ziqi Xu (CSE)

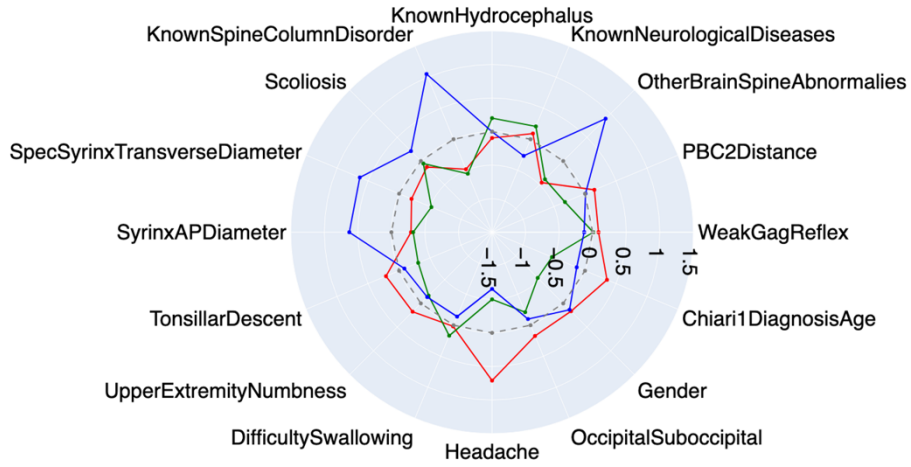
Clustering

- **Identified three distinct clusters**
- **Reproducible across internal/external cohorts**
- **Inform clinical decision-making and management**

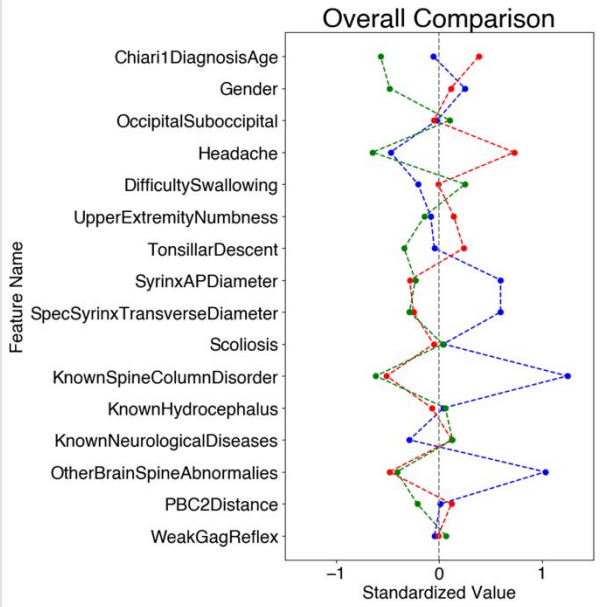
Gupta VP, Xu Z, Greenberg JK, Strahle JM, Haller G, Meehan T, Roberts A, Limbrick DD Jr, Lu C. Using artificial intelligence to identify three presenting phenotypes of Chiari type-1 malformation and syringomyelia. *Neurosurgery*. 2024 Nov 18.

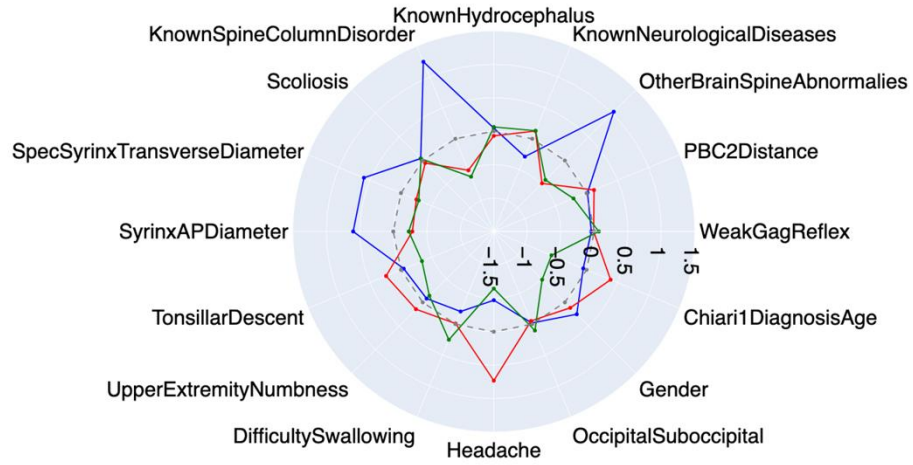
Curve and Radar Plot for Internal Cohort





Curve and Radar Plot for External Cohort





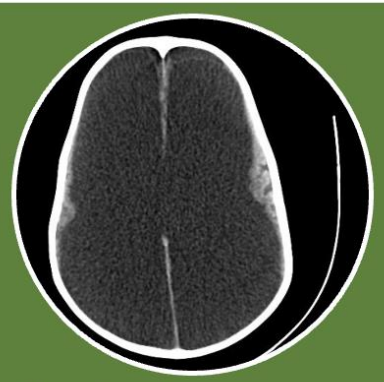
Clinical Implications

Classic CM1 + syrinx
- Headache-dominant
- Few additional symptoms

Neonatal comorbidities
- Few classic CM1 symptoms
- Progression under study

Syrinx-related spinal effects
- Early diagnosis → may prevent major surgery

Cluster 1
- Highest diagnosis age
- More likely female
- Presents with HA
- Fewer bulbar symptoms
- Greatest tonsillar descent
- Small syrinxes
- Few other medical issues
- n=277

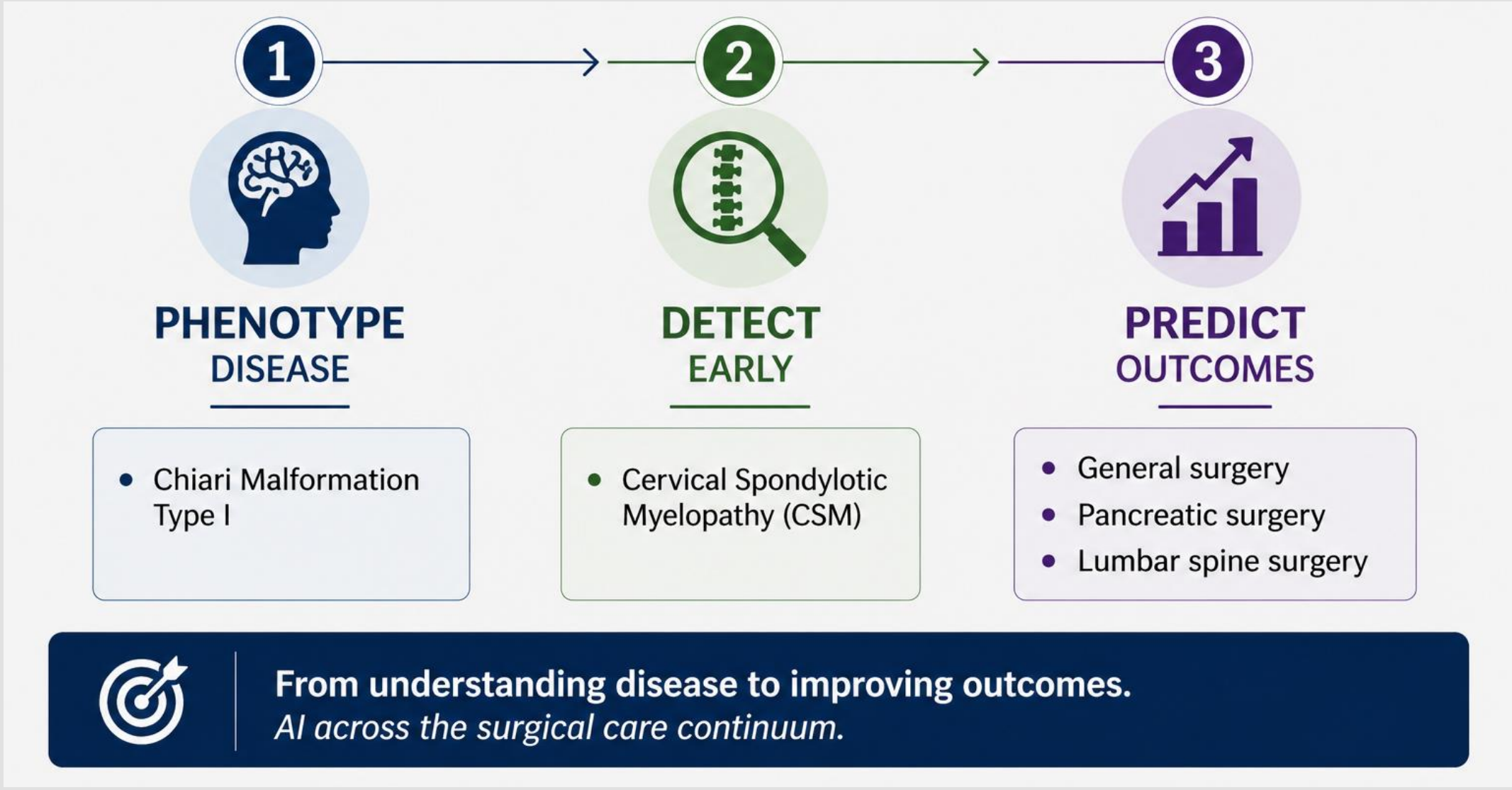


Cluster 2
- Youngest diagnosis age
- Less tonsillar ectopia
- Small syrinxes
- Many newborn issues
- Fewer headaches
- Most hydrocephalus
- Most bulbar symptoms
- n=258

Cluster 3
- Less tonsillar ectopia
- Largest syrinxes
- Most spine disorders
- Most scoliosis
- Most motor deficits
- Fewer headaches
- Fewer bulbar symptoms
- n=187

- Typical radiologic definition is inadequate for heterogeneity of CM1
- Identified clusters potentially assist clinicians direct clinical management

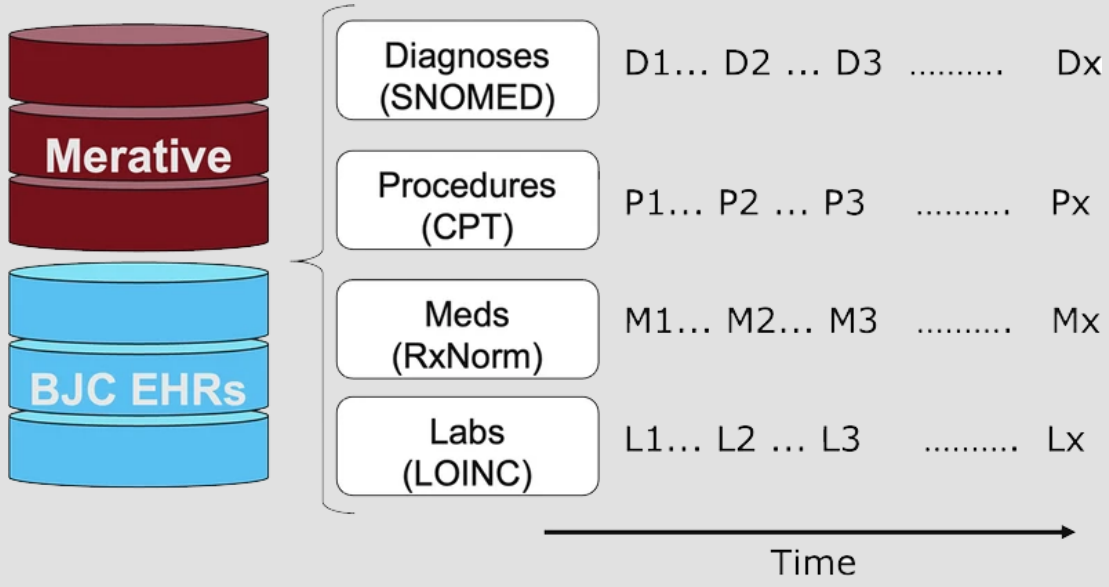
AI for Surgery: Roadmap



Predict Cervical Spondylotic Myelopathy (CSM)

- CSM is the **leading cause** of spinal cord dysfunction in older adults
- Diagnosis is frequently delayed by **2–6 years** after symptom onset
- Goal: Predict CSM using **longitudinal EHR data** before clinical diagnosis

Data Sources
- Merative MarketScan: ~1.4M patients' longitudinal claims data
- BJC Health: ~500K patients' institutional EHRs



Work with Jacob Greenberg, Salim Yakdan (Neurosurgery), Ben Warner (CSE)

Models for Structured EHR Data

- **Foundation models**
  - **`clmbr-t-base`**: **141M** parameters, pretrained on 2.57M Stanford Hospital patients.
  - **`clmbr-t-5k-csm`**: **145M**-parameter model with an additional 5K medical codes frequently associated with CSM.

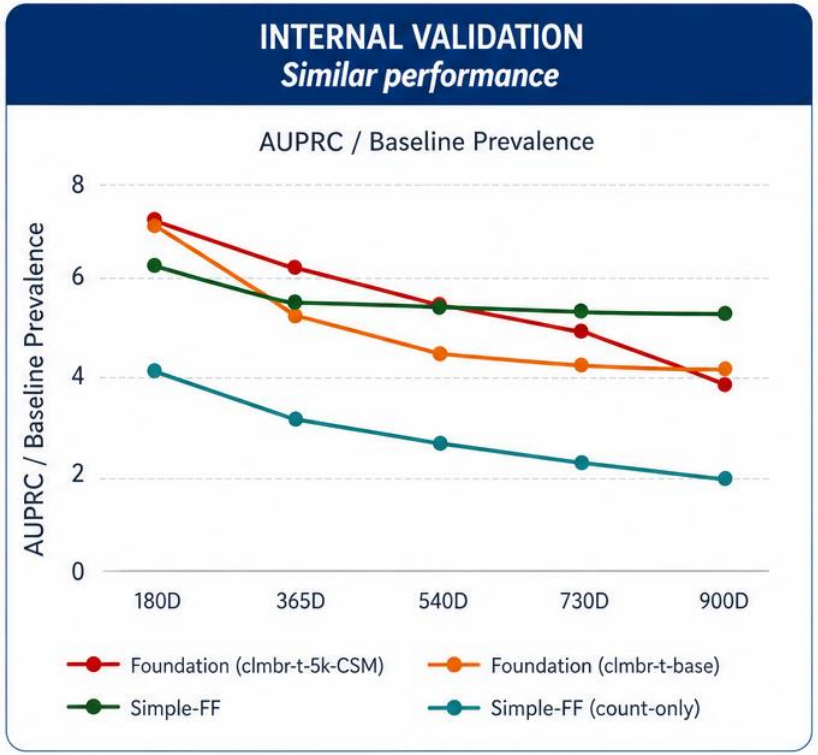
- **Clinically guided models**
  - **`simple-ff-count-only`**: **391K** parameter 3-layer neural network measuring the counts of 497 codes selected by **clinicians**.
  - **`simple-ff`**: **399K** parameter 3-layer neural network measuring the same 497 codes plus selected patient characteristics.

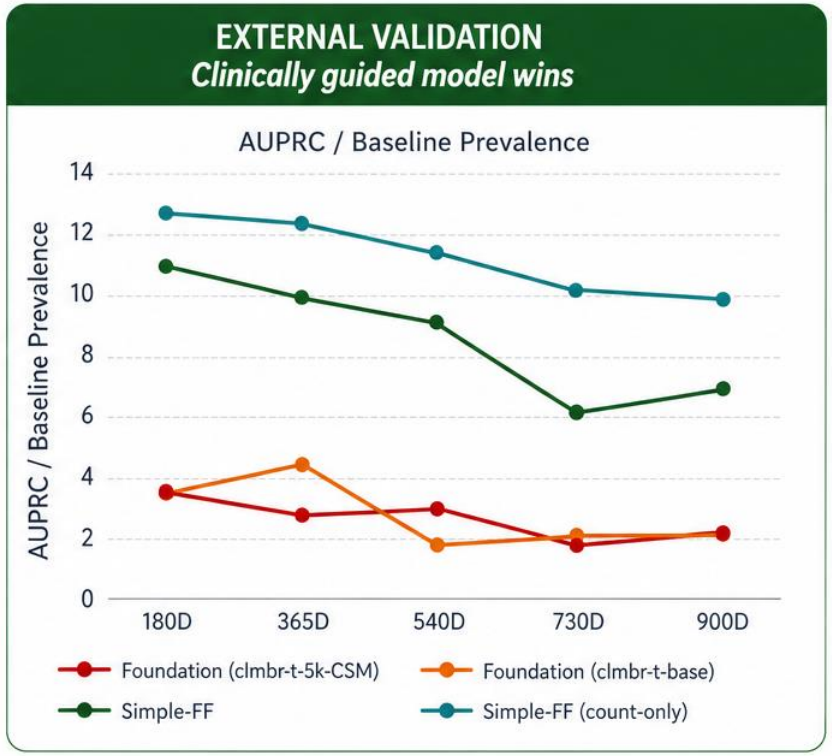
Performance & Generalizability

INTERNAL VALIDATION:
Similar performance across models

EXTERNAL VALIDATION:
Clinically guided model outperforms

KEY TAKEAWAY:
Clinical knowledge drives better generalization

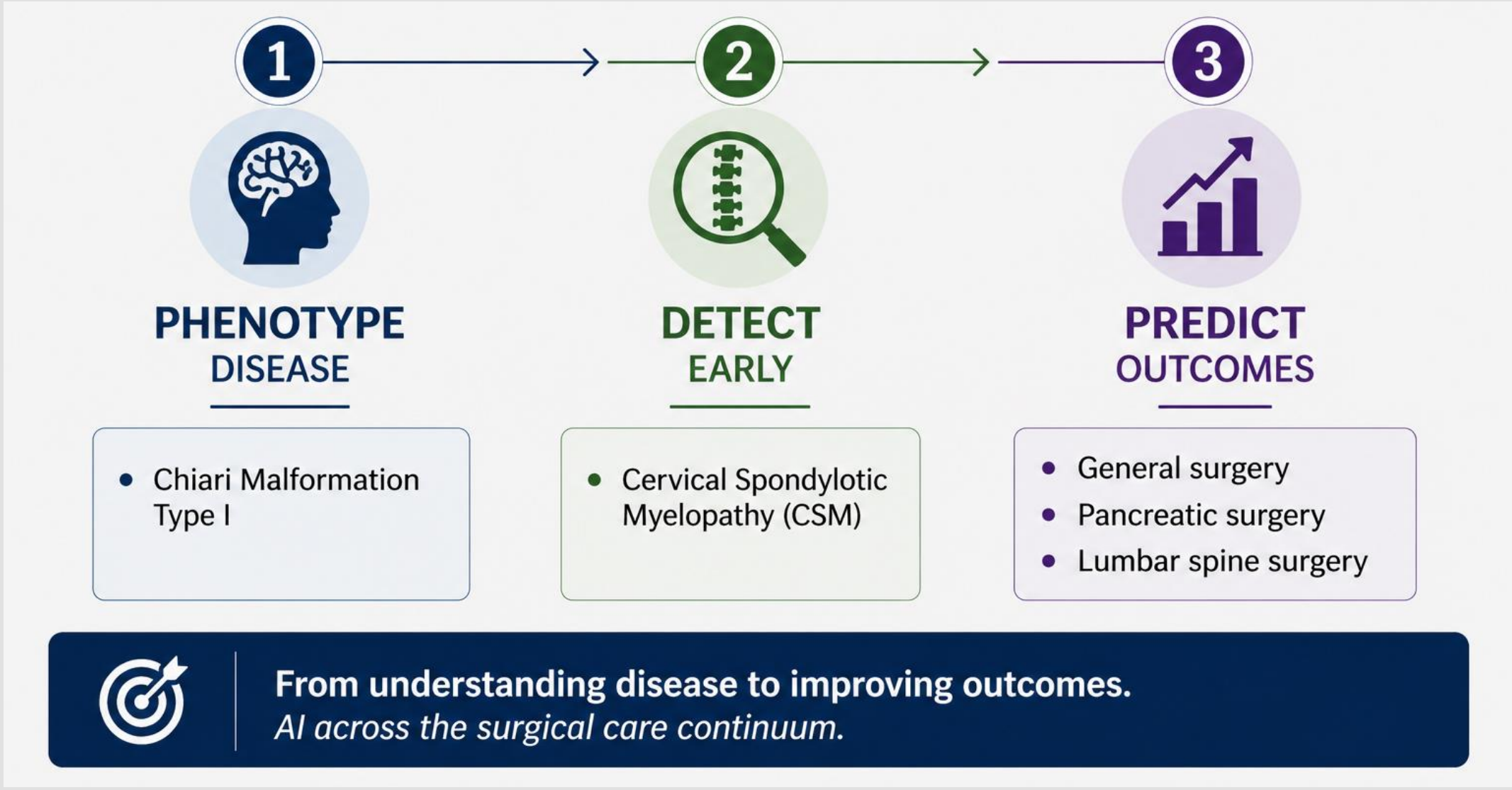




CLINICAL KNOWLEDGE → BETTER GENERALIZATION → BETTER PATIENT CARE

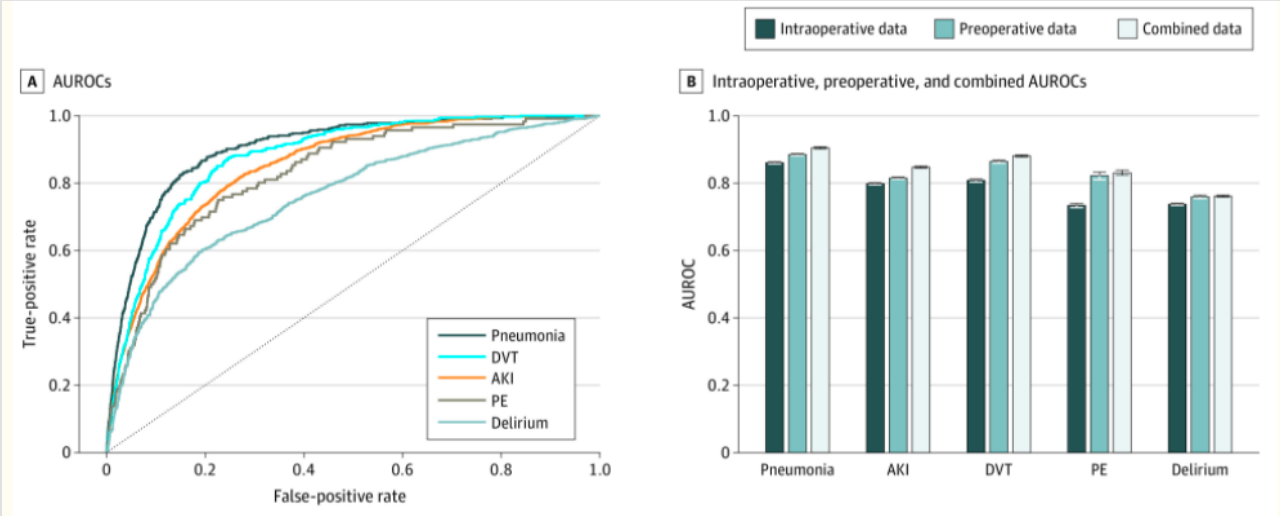
Yakdan S, Warner BC, Ghogawala Z, Ray WZ, Bydon M, Steinmetz MP, Griffey RT, Foraker R, Wilcox A, Lu C, Greenberg JK. Clinically-guided models or foundation models? Predicting cervical spondylotic myelopathy from electronic health records. *npj Digital Medicine*. 2026;9:153.

AI for Surgery: Roadmap



Predict Postoperative Complications

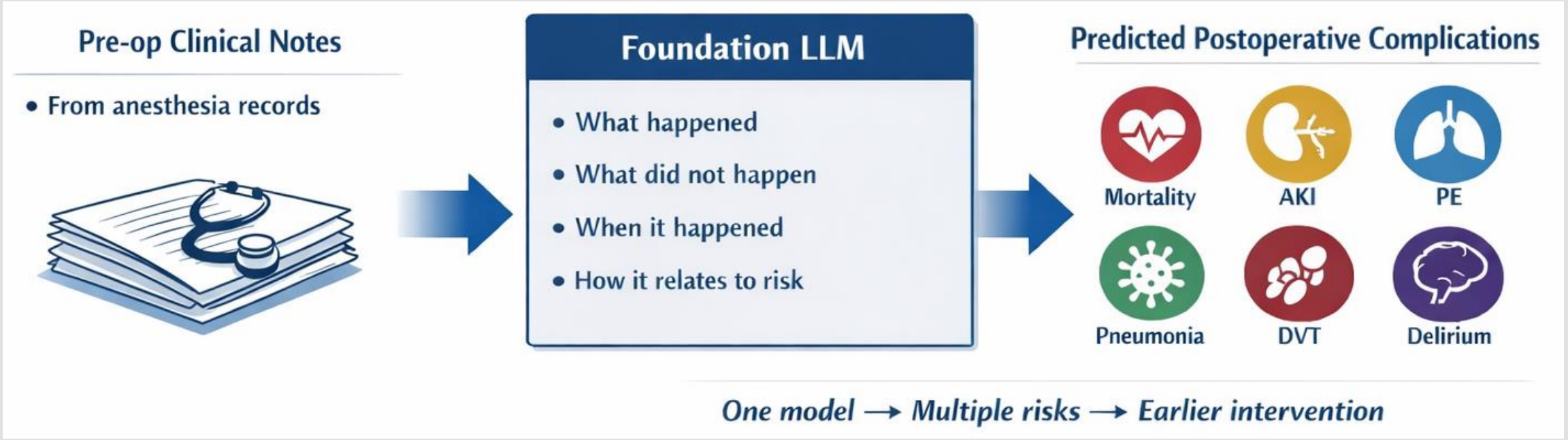
- Benefits of **predictive** models
  - Preemptive and early identification of risk factors
  - Anticipatory management for perioperative contingency planning
  - Shortened hospital stays and reduced complication risks
- Predictive models use **structured** EHR data



B. Xue, D. Li, C. Lu, C.R. King, T. Wildes, M.S. Avidan, T. Kannampallil, and J. Abraham, Use of Machine Learning to Develop and Evaluate Models Using Preoperative and Intraoperative Data to Identify Risks of Postoperative Complications, *JAMA Network Open*, 4(3):e212240, 2021.

Predict Postoperative Risks from Clinical Notes

- Clinical notes capture nuanced clinical judgment that structured data cannot.
- **Large Language Models (LLMs)** translate narrative documentation into risk signals.

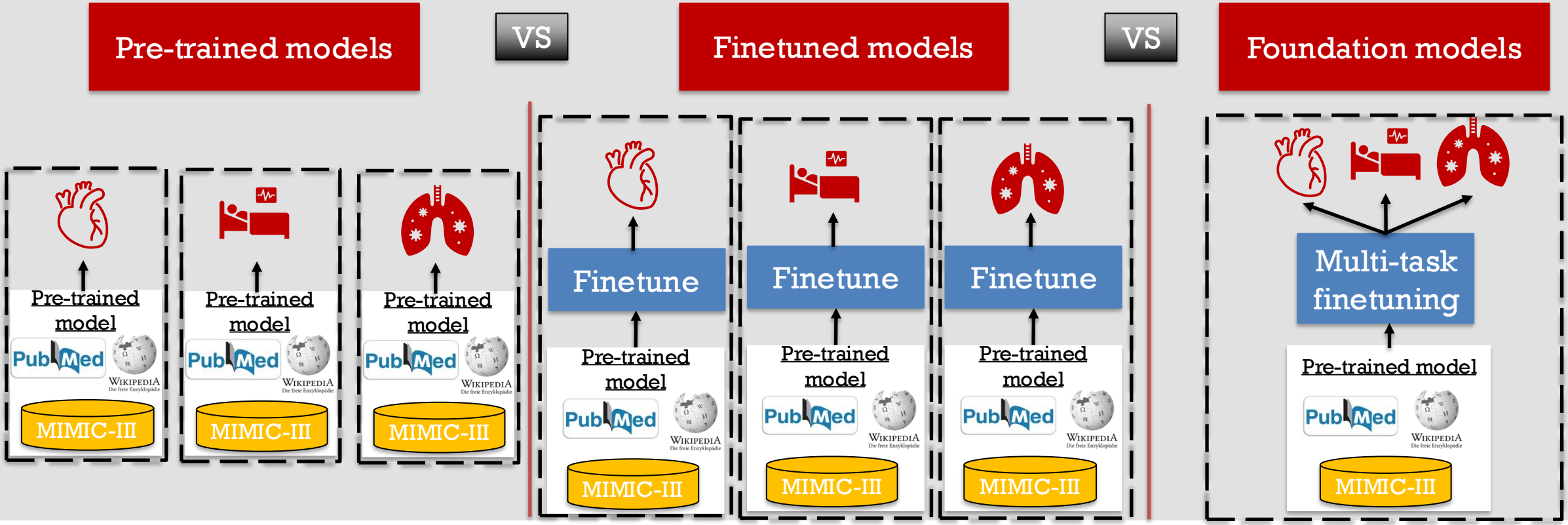


Dataset: 84,875 pre-operative anesthesia notes from Barnes-Jewish Hospital (BJH)

Work with Joanna Abraham, Thomas Kannampallil (Anesthesiology), Charles Alba (DCDS)

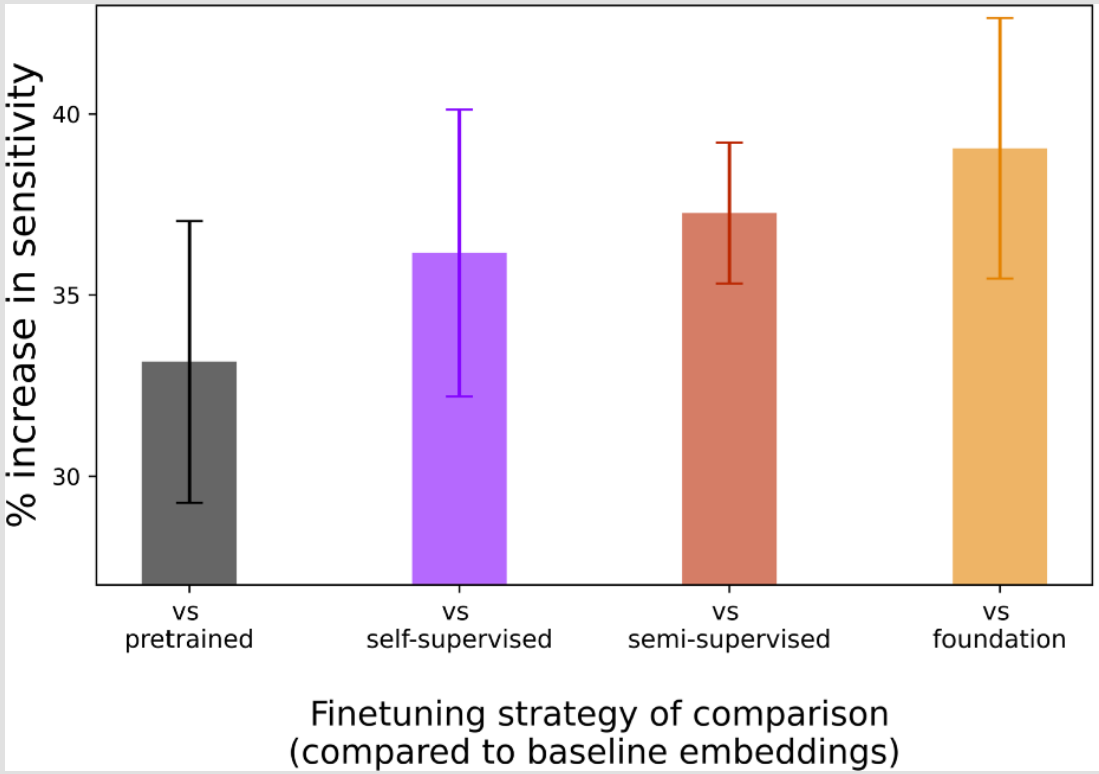
LLMs for Predicting Postoperative Risks

1. Do **pre-trained** LLMs outperform traditional NLP?
2. Does **fine-tuning** to surgical notes improve performance?
3. Can a single **foundation** model predict multiple risks?



Foundation LLMs Improve Surgical Risk Prediction

- Pretrained: **+33** percentage point increase over NLP in sensitivity (at 95% specificity)
- **One** foundation model predicts **multiple** complications
  - **Outperforms** individually fine-tuned models
  - **Simplifies** clinical deployment

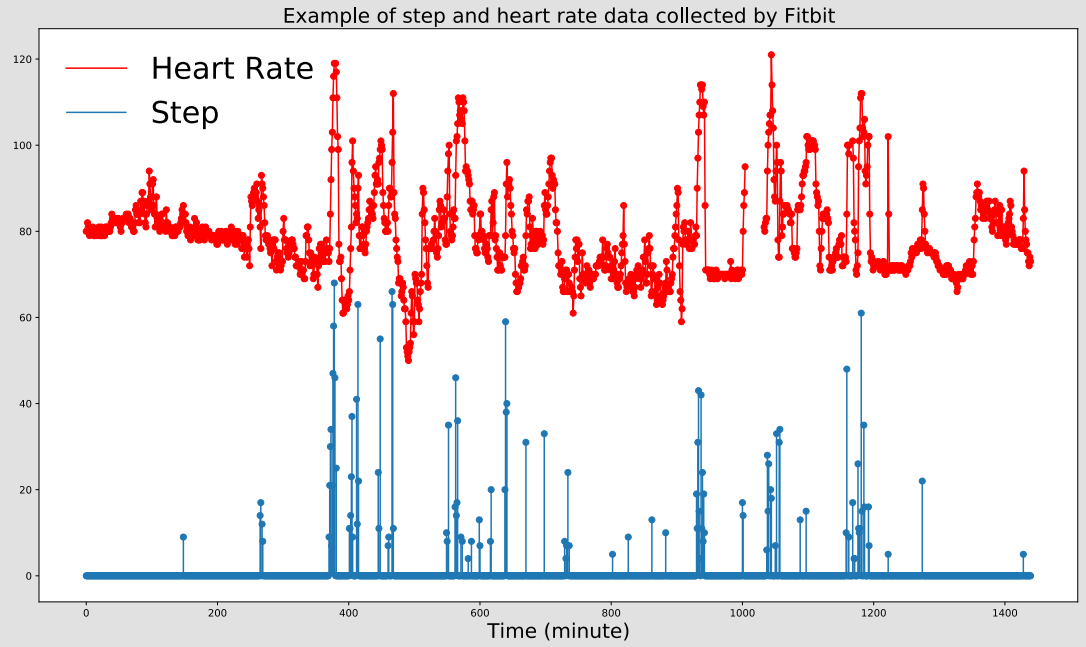


The number of high-risk patients correctly identified per 100 cases that were missed by baseline models.

Alba C, Xue B, Abraham J, Kannampallil T, Lu C. The foundational capabilities of large language models in predicting postoperative risks using clinical notes. *npj Digital Medicine*. 8, 95 (2025).

Wearables

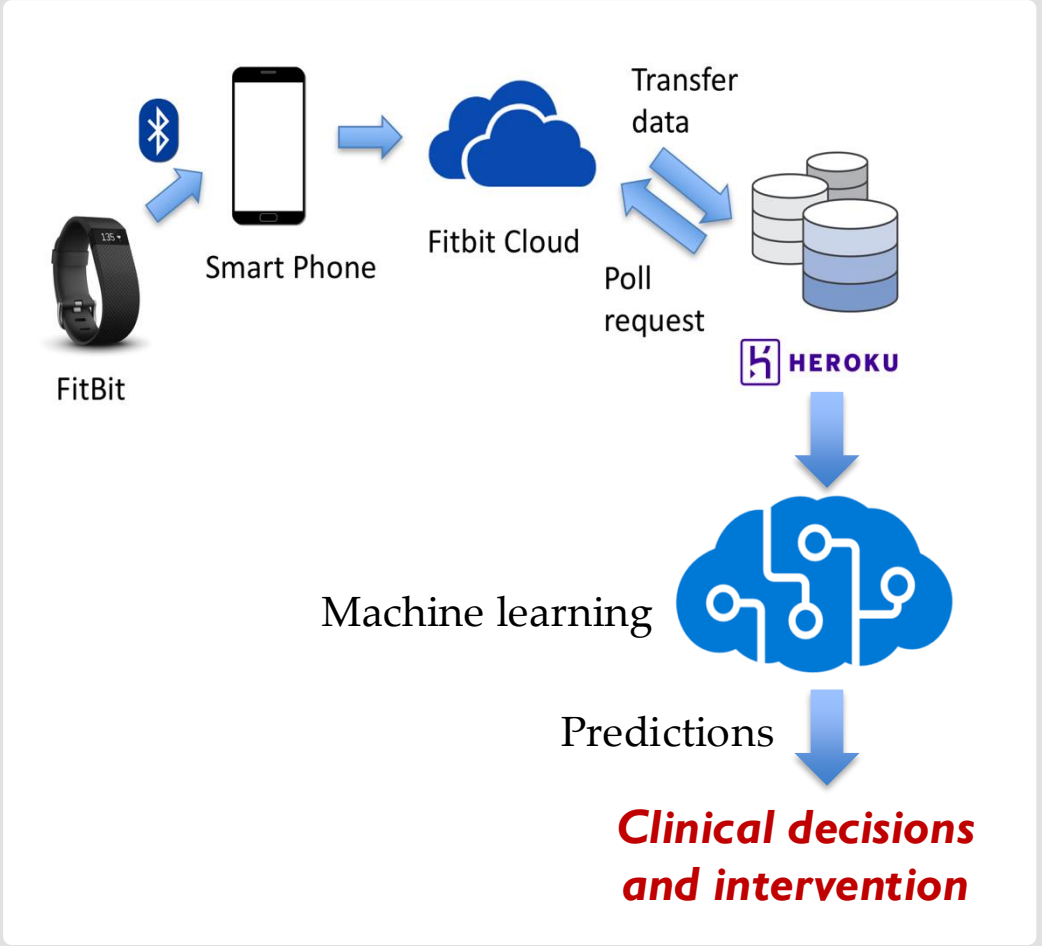
- Commonly available: step, heart rate, sleep stages
- More sensing modalities
  - Oxygen saturation (SpO2)
  - Skin temperature
  - Breathing rate
  - Heart rate variability
  - ECG
  - Stress
- **500+ million** wearables sold in 2021



Unobtrusive, longitudinal monitoring outside hospitals!

Decision Support for Pancreatic Surgery

- Pancreatic cancer has a 5-year survival rate less than 5%.
- Surgery is commonly followed by complications.
  - Pancreatectomy: 90-day mortality rate over 4% and serious morbidity rates over 20%.
- Predict postoperative complications before surgery
  - **Decision support**: suitability for surgery
  - **Intervention**: prehabilitation



Work with Chet Hammill (Surgery), Jingwen Zhang (CSE)

Robust Feature Engineering Pipeline

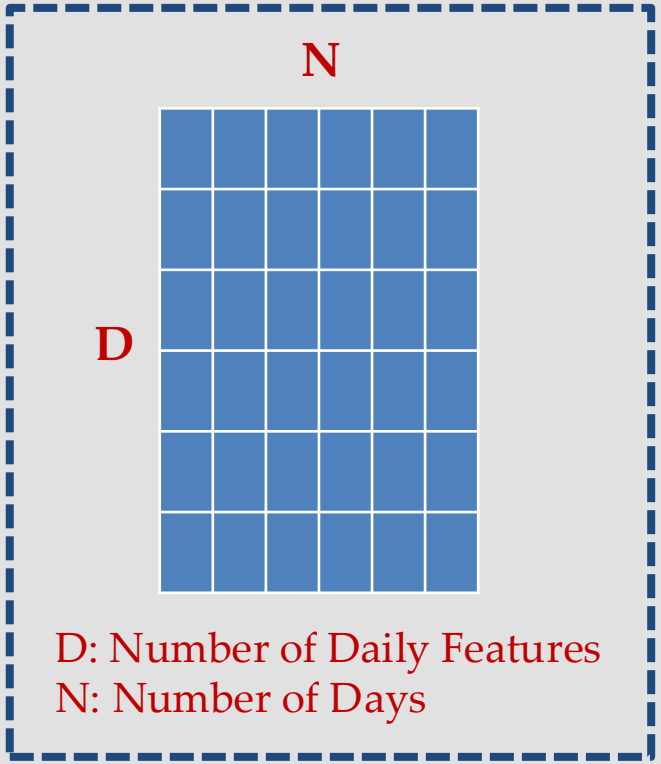
Daily Feature Extraction

Step Count

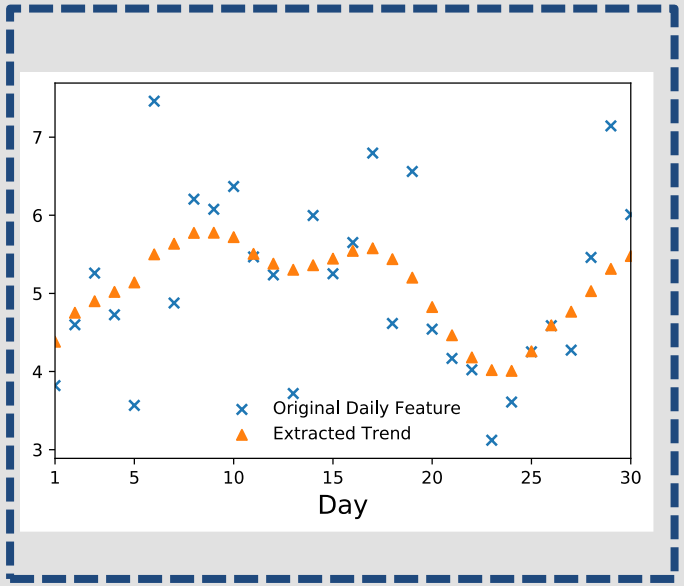
Heart Rate

Sleep Stage

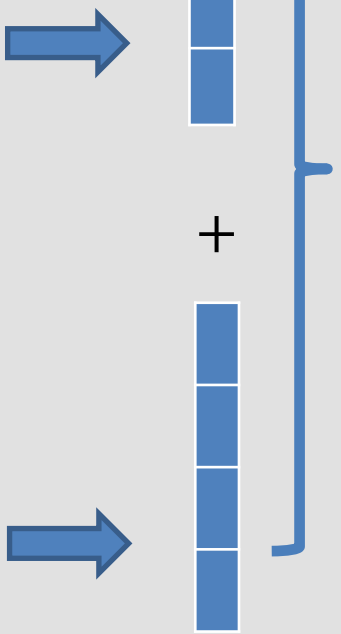
Wearable Time Series



High-level Feature Extraction



Clinical Characteristics



Predict Complications after Pancreatic Surgery

Machine learning outperforms 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

Expanding to the Surgical Prehabilitation and Readiness (SPAR) program

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*, 6(2), Article 87, July 2022.

Predicting Lumbar Spine Surgical Outcomes

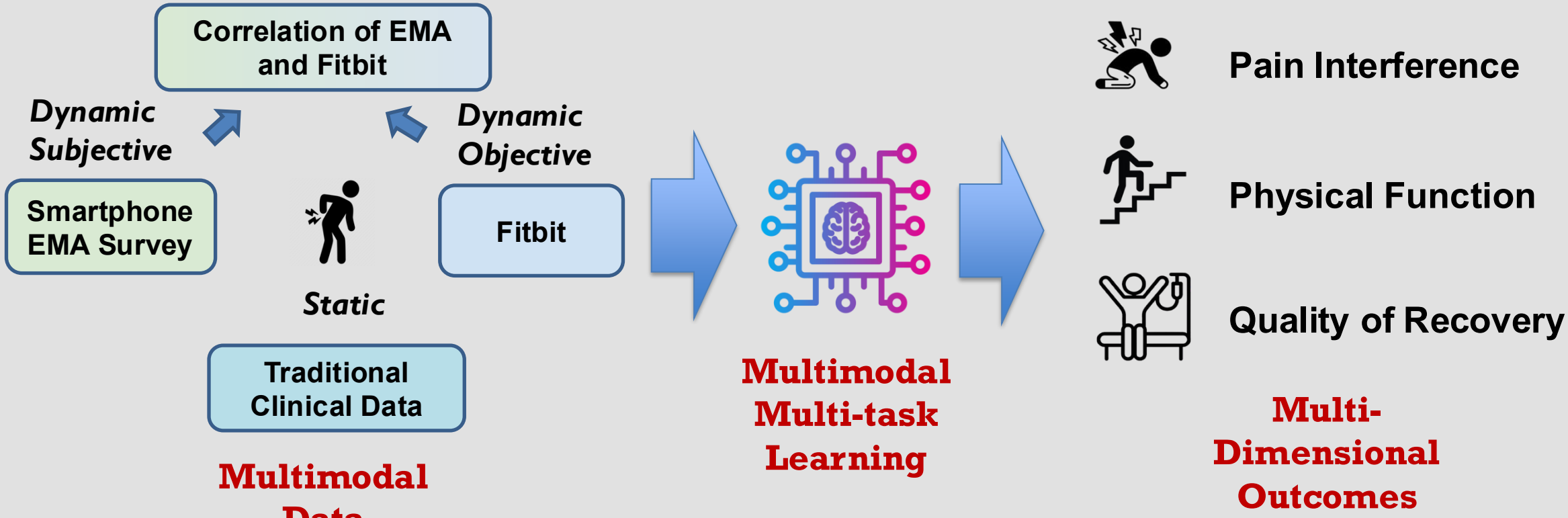
- Burden of low-back chronic pain
  - Affects 1.36 million people and cost $135 billion in USA
  - Lumbar spine surgery is potential treatment

- **Heterogeneous** and **multi-dimensional** surgical recovery
  - Interrelated: **pain, physical function, mental health**

Goal: Predict **multi-dimensional** recovery using **multi-modal** data

Work with Jacob Greenberg (Neurosurgery), Ziqi Xu (CSE)

M³L: Multi-Modal Multi-task Learning



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).

Clinical Study

- 122 lumbar spine surgery patients at a large academic medical center

- **RMSE**: Root Mean Square Error
- **Lower** RMSE → better prediction performance

	Δ Pinter (Range 43.4)	Δ PhysFun (Range 38.1)	QOR (Range 102)
	RMSE	RMSE	RMSE
Best Single-Task	7.6903	6.6969	18.1399
Traditional Multi-Task	8.3126	6.6828	17.4709
Our Model	**7.4462**	**6.4245**	**14.9226**

- **Single-task** models miss complex recovery associations
- **Traditional multi-task** → negative transfer
- **Our model** → consistently better + more robust

Predicting Postsurgical Persistent Pain

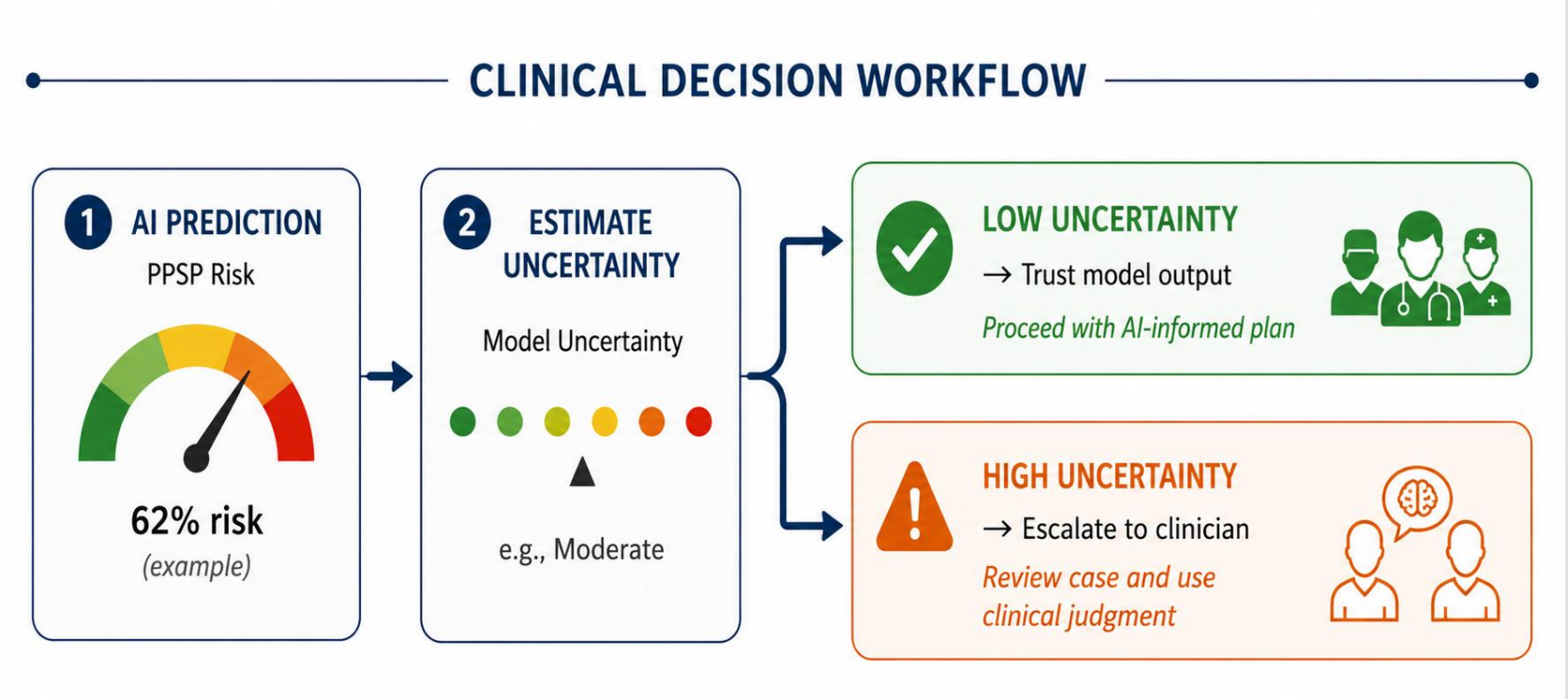
- Persistent Postsurgical Pain (PPSP)
  - Pain lasting >3 months after surgery
  - Affects 10–35% of patients (among the most common complications)
  - Impacts ~45 million patients/year with substantial clinical and economic burden
- Limitations of Current Approaches
  - Rely primarily on static clinical data
  - Miss dynamic patient factors (e.g., mental health, activities)
  - No uncertainty estimate → limits clinical trust

Goal: Multimodal, uncertainty-aware AI for PPSP prediction

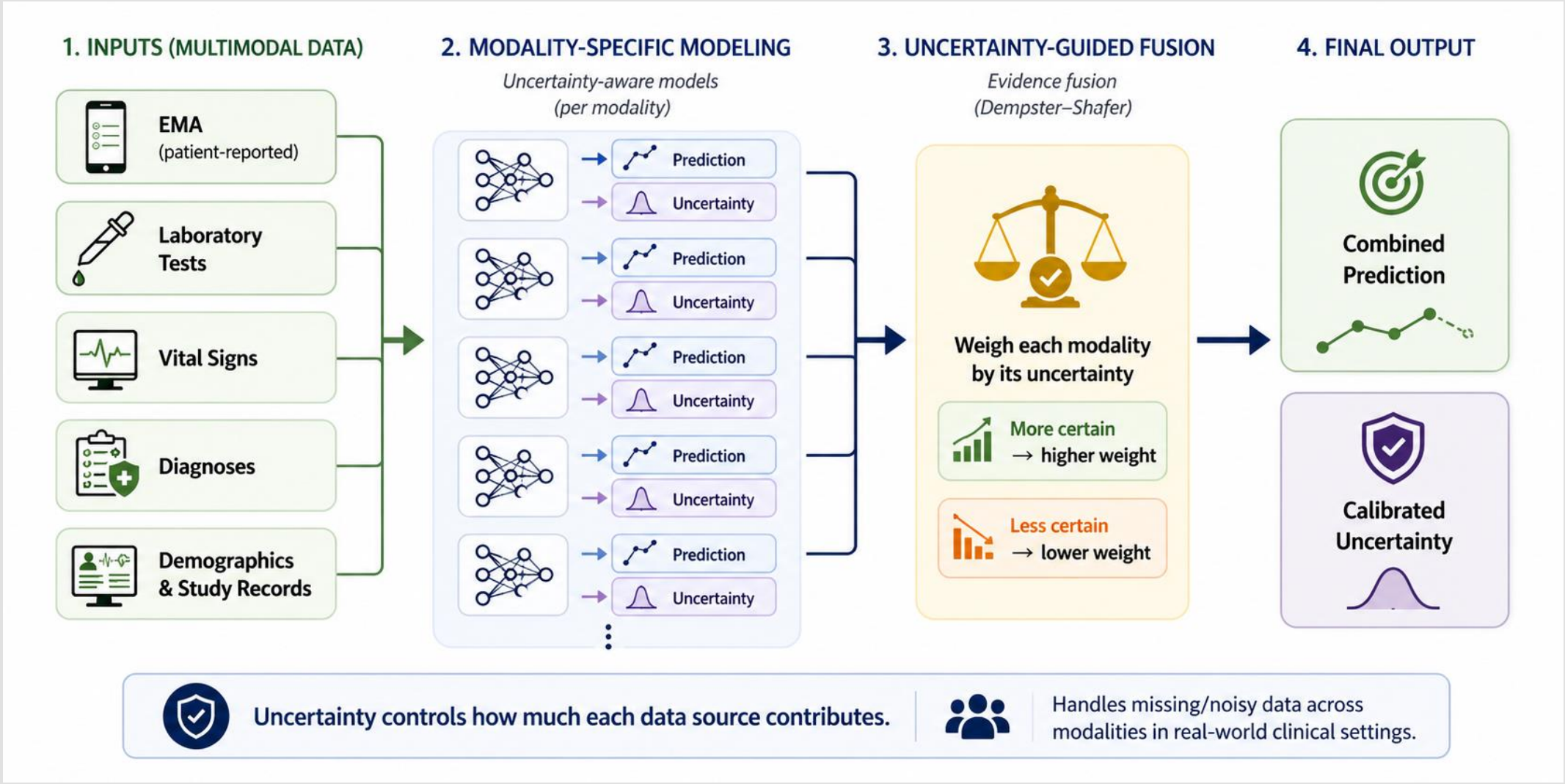
Joint work with Simon Haroutounian (Anesthesiology), Ziqi Xu (CSE)

Why Uncertainty Matters in Surgery

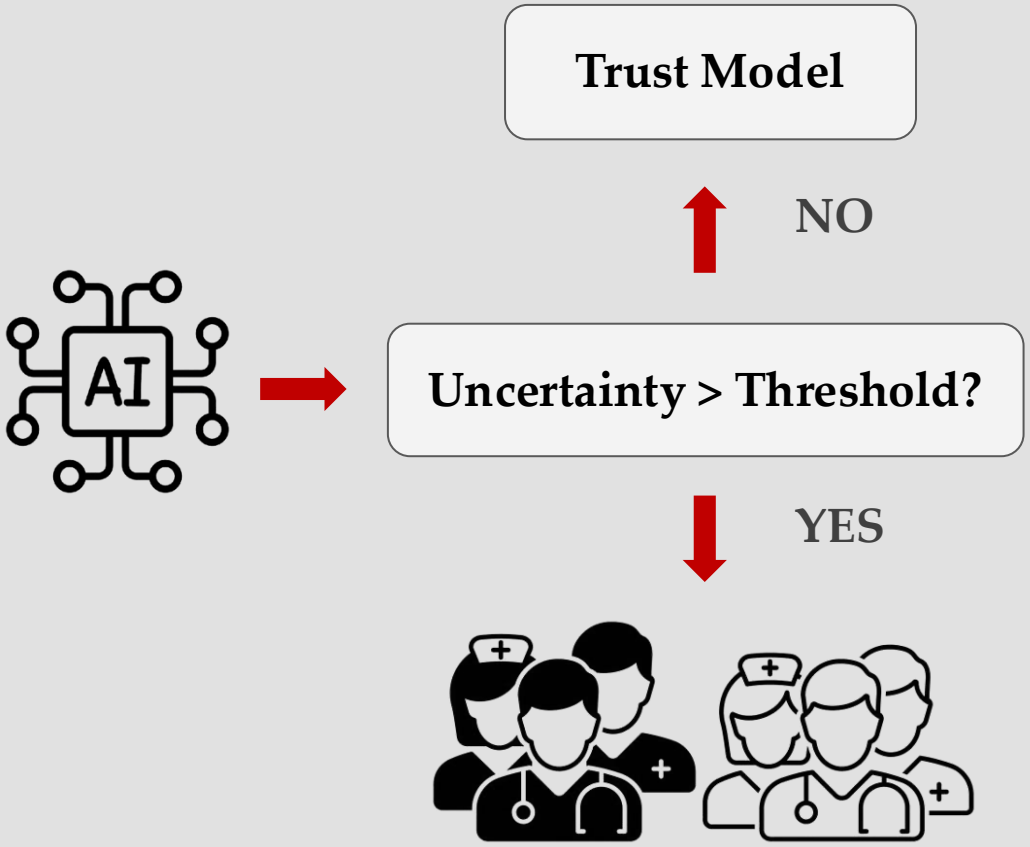
- **High-stakes** decisions: complications, recovery, long-term outcomes
- AI can be **confidently wrong** → risk of misleading decisions
- **Uncertainty** determines when to trust AI and when to defer to clinicians

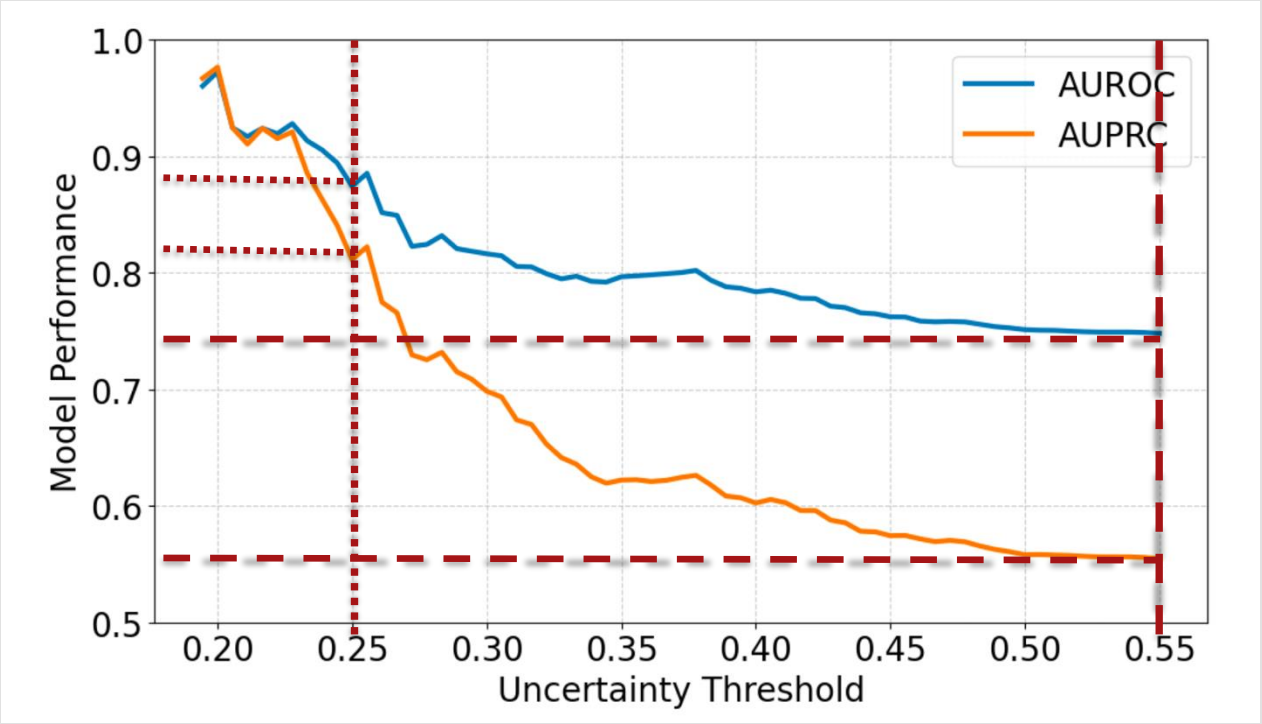


Uncertainty-Aware Multimodal AI



Predictive Performance vs. Uncertainty





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

AI is Transforming Surgical Care

- Phenotyping patients → precision interventions
- Early detection → early treatment
- Outcome prediction → personalized care
- Uncertainty-aware AI → trustworthy decisions

AI for Surgery: Looking Ahead

- **Human–AI Collaboration**
  - Explainable
  - Uncertainty-aware
  - Integrated into clinical workflow
- **Robust Models**
  - Generalize across sites
  - Remain reliable over time
- **Stronger Partnerships**
  - AI experts × surgeons
  - Co-design solutions that matter

Bring AI and Health Experts Together to Tackle Critical Challenges

https://aihealth.wustl.edu/