# STAT 992: Foundation Models for Biomedical Data

Ben Lengerich

Lecture 12: Foundation Models for Clinical Images

March 9, 2026

# Logistics – Review Paper

- Due dates:
  - Write first draft of your section by **April 13th**
  - Review 1 section by **April 20th**
  - Update your section by **April 29th**

# Today: Imaging

- Classic ML Pipeline:
  - Dataset → Training → Model → Prediction
- Foundation Model Pipeline:
  - Pretrained Model → Prompt → Prediction

Modern foundation models adapt primarily **at inference time rather than training time.**

# Medicine generates enormous imaging data

- Examples:
  - Billions of radiology scans annually
  - Pathology slides digitized at gigapixel scale
- Major categories:
  - Radiology
  - Pathology
  - Dermatology
  - Ophthalmology
  - Ultrasound

**But annotation is expensive.**

# The supervision bottleneck

**Annotation is expensive.**

- Labeling requires specialists:
  - Tumor segmentation → Pathologist
  - Fracture detection → Radiologist
  - Lesion grading → Dermatologist
- Foundation models shift the pipeline:
  - Traditional:
    - Labeled dataset → Train task model
  - Foundation models:
    - Massive unlabeled images → Representation learning

    → Small labelled dataset → Adaptation

# The reuse question

- Two paradigms:

| Foundation Model Paradigm | Domain-Specific Paradigm |
|---|---|
| Large shared representation → fine-tune | Specialized dataset → Specialized model |
| Reuse across tasks | Tailored to local workflows |
| Less Labeled Data | Fewer distribution shifts |
| Rapid Deployment | Simpler models |

Key question: Is structure actually shared across hospitals and imaging modalities?

# Medical imaging challenges

- **Scale** - Pathology whole-slide images can easily be 100,000 x 100,000
- **Weak supervision** – Often labels only exist at higher granularity (image, slide, or patient level)
- **Distribution shift** – Medical imaging varies across hospitals. Models trained at one hospital tend to fail at another.
- **Shortcut learning** – Model predicts outcome because of associated signal (e.g. hospital identifiers, demographic correlations, portable X-ray markers)

# Two Imaging Regimes

# Radiology vs Pathology

- Clinical imaging splits roughly into two regimes:
  - Radiology
    - X-ray, CT, MRI
  - Pathology
    - Microscopic tissue images
- Very different from a systems perspective

# Radiology Scale

- Typical chest X-ray: 1024 x 1024 pixels
- Models can process the whole image directly
- Standard CNN or ViT works

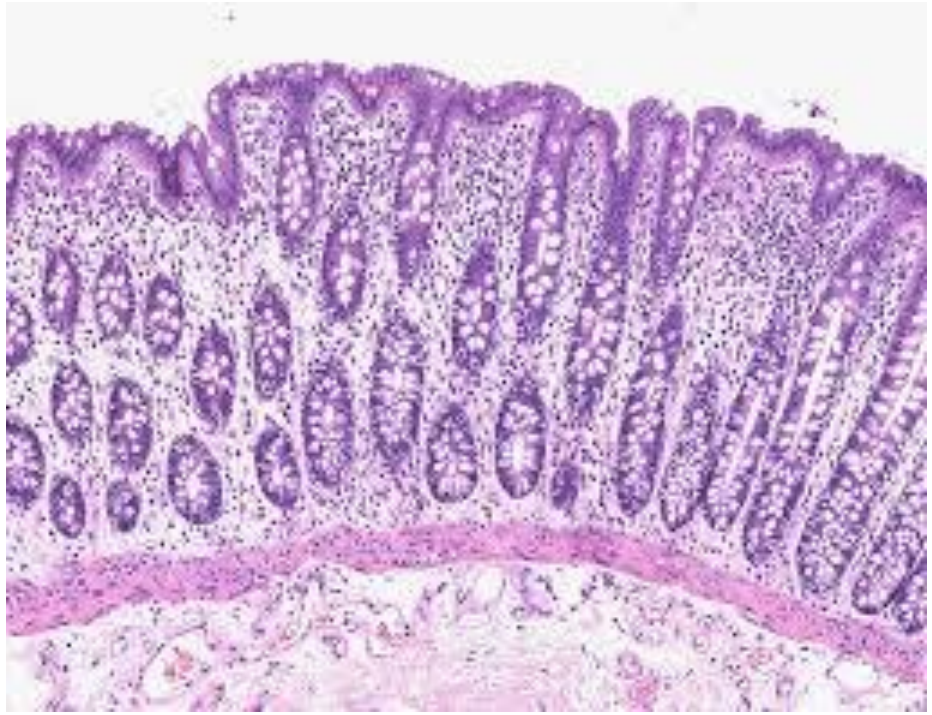
- A unique advantage: radiology images are often paired with reports

# Pathology Scale

- Typical whole slide image: 100,000 x 100,000 pixels
- Impossible to feed directly to a neural network (?)

- Typical pipeline: Tiling
- Slide → Tiles → Tile embeddings → Aggregation → Prediction

- Pathology models rarely see the full slide; must reason indirectly through tile statistics.



# Pre-training strategies

# Three major strategies

- Foundation models for images use:

| Method | How | What the model learns |
|---|---|---|
| Contrastive Learning | align embeddings of related pairs | semantic concepts shared across images and text |
| Masked Image Modeling | reconstruct masked patches | spatial structure and visual features |
| Self-supervised Tile Learning | learn from tile relationships | cellular or tissue-level morphology |

# Contrastive Learning

- CLIP-style objective

$$\max \log \frac{\exp\left(\frac{sim(I,T)}{\tau}\right)}{\sum_T' \exp\left(\frac{sim(I,T')}{\tau}\right)}$$

where $I$ is image embedding, $T$ is text embedding.

- Used heavily in radiology.

Should imaging foundation models learn to include the information typically encoded in text?

# Masked Image Modeling

- Train a model to predict missing parts of an image from context.
- Training procedure:
  - Split image into patches
  - Randomly mask a subset of patches
  - Encode the visible patches
  - Predict the missing patches

Objective: Learn representations that capture **spatial structure and semantics**.

# Self-Supervised Tile Learning

- Problem: Whole slide images are extremely large.
- Training procedure:
  - whole slide
  - divide into tiles
  - encode each tile
  - learn representations from tile relationships

Learn cell and tissue representations without slide-level labels.

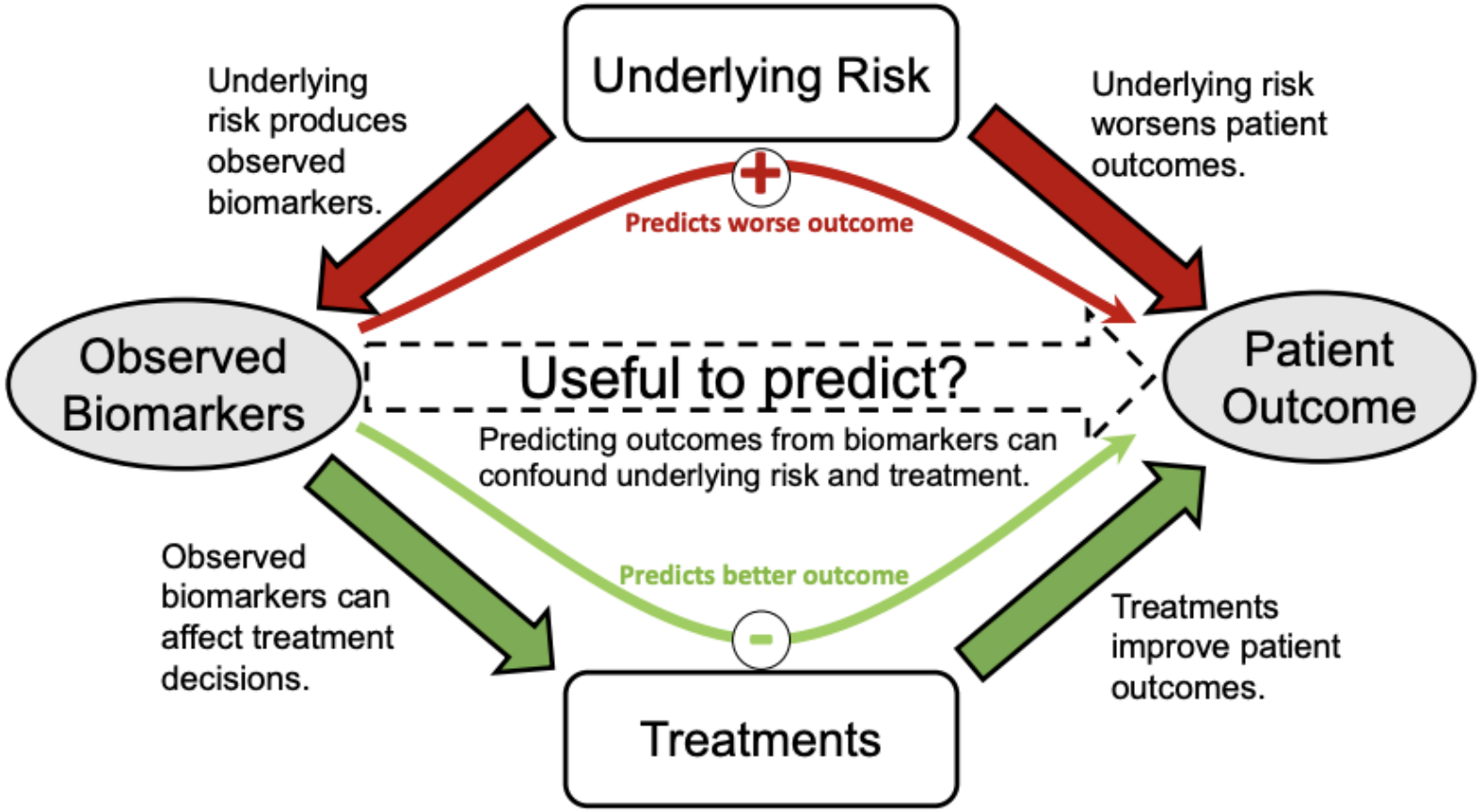
# In Medical Context

# Many downstream tasks for foundation models

- A single foundation representation can support:
  - disease classification
  - report retrieval
  - abnormality detection
  - image search
  - triage systems

# The causal problem

- Medical images are not purely biological



# **The causal problem** in practice

- Example: portable X-rays
- Portable scanners are more frequently used for **sicker patients**.
- So the dataset contains associations like:
  - portable scanner marker → pneumonia
- But the marker is not biological.

# **The causal problem:** representation learning dilemma

**Should the representation encode signals caused by clinical workflow?**

- Re-use strategies assume “yes”. Assume there are downstream “de-confounding procedures” to extract types of signal.

# Benchmarks often mislead

- Most benchmarks evaluate:
  - f(image) → diagnosis
- But:
  - Recorded “diagnoses” are often billing codes
  - Real clinical decisions require context
    - Example: Chest X-ray shows lung opacity
    - Possible causes: pneumonia, pulmonary edema, tumor, atelectasis
    - Clinicians resolve this using: symptoms, labs, prior scans

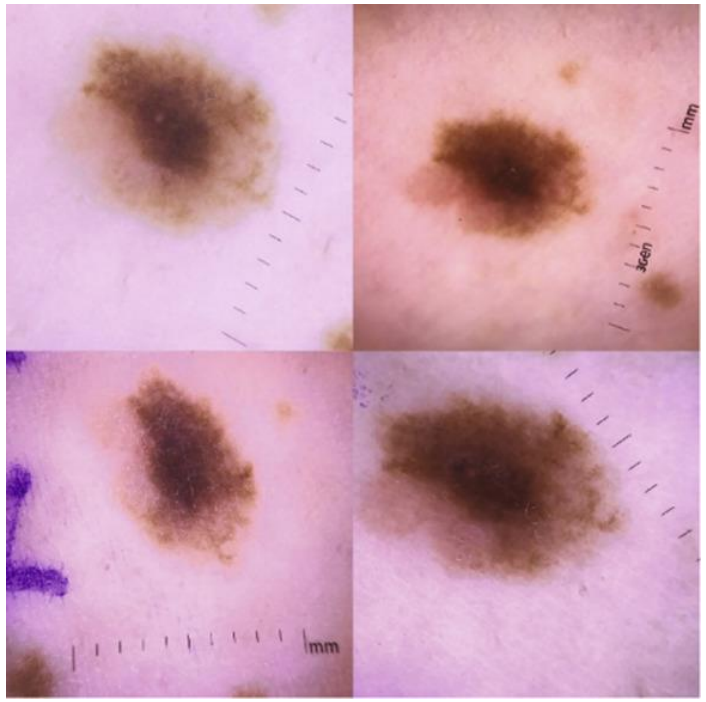
# Label Noise

- Radiology labels often come from reports.
- Reports contain:
  - visual observations
  - clinical hypotheses
  - prior knowledge
- Example: “Possible pneumonia given symptoms”

**Can we remove signals caused by clinical workflow from our learned representations?**

# Shortcut learning

- Models often rely on spurious signals
- Example:
  - scanner artifacts
  - hospital identifiers
  - rulers in dermatology images



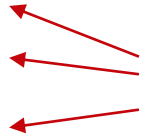
# What actually transfers?

- Foundation models rely on shared structure.
- Possible invariances:
  - anatomical structure
  - cell morphology
  - physics of imaging
- If these transfer, reuse works.
- If not, domain-specific models dominate.

May be different conclusions for radiology vs pathology.

# The big picture

- Foundation models change:
  - Representation reuse
  - Training from unlabeled archives
  - Integration with multimodal systems
- Foundation models do not automatically solve:
  - Distribution shift
  - Shortcut learning
  - Clinical reasoning



Robust statistical signals

The central research question: **What invariances should biomedical foundation models learn?**

# Questions?