Six Ideas in AI Evaluation
First Term 2026, Johns Hopkins Bloomberg School of Public Health

Instructor: Yiqun T. Chen
Office: Wolfe Street Building, E3608
Office Hours: Schedule via email
E-mail: yiqunc@jhu.edu

Course Meeting Times: Mondays 3:30–4:50 pm
Location: TBD
Website: TBD

Course Description:
Imagine your manager or collaborator pitches you an AI-based system — whether for ad targeting or for adverse event prediction — and asks how it should be deployed and evaluated. You have a budget, some data, and a deadline. What study do you run, and what evidence would make you say go or no-go?

This course is about starting to answer that question. We will not cover the details of any AI method, but rather their high-level properties — stochasticity, bias, a fast product life cycle — and how those properties bear on the design of an evaluation. The six topics are prediction evaluation, benchmarks, model-generated labels, distribution shift, study design, and subgroup performance. Each is taught on small semi-synthetic datasets where the truth is known, so that failure modes can be seen directly rather than described.

Intended Audience: Master's and PhD students in Biostatistics or other departments with quantitative training, or by instructor permission.

Prerequisites: Familiarity with regression and basic inference. Students should have used an LLM to solve a technical problem, broadly defined. Some programming experience in Python or R, or a willingness to learn — including by vibe coding. No prior exposure to deep learning is assumed.

Evaluation and Grading: Pass/Fail. Three short assignments, due after Weeks 2, 4, and 6.

Learning Objectives:
Upon completion of this course, students will be able to:

- Evaluate a prediction system.
- Judge what a benchmark score does and does not imply about a deployment.
- Draw valid conclusions when outcomes are labeled by a model (e.g., an LLM) rather than by an expert gold standard.
- Diagnose whether and why performance degrades across sites and populations.

- Design a validation study with an explicit sampling plan and budget.
- Assess subgroup performance and the trade-offs among fairness criteria.

Course Materials: Slides, notebooks, datasets, and recommended readings will be posted on the course website.

Weekly Schedule (Tentative)

- *Week 1 (Aug 24):* **Evaluating predictions.** Accuracy and calibration; classification and calibration metrics; proper scoring rules; and the behavior of these metrics under low prevalence and class imbalance.
- *Week 2 (Aug 31):* **Benchmarks.** What is a benchmark, and what is it measuring? Construct validity, contamination, leaderboards, and Goodhart's law. Steps for building an evaluation set for a system nobody has benchmarked.
- *(Sep 7): No class — Labor Day.*
- *Week 3 (Sep 14):* **Model-generated labels.** Human labels are expensive, so AI models are increasingly used to augment them. Measurement error, prediction-powered inference, and the reliability of LLM annotators.
- *Week 4 (Sep 21):* **Distribution shift.** Performance may shift across time, sites, and subgroups. How do we diagnose this empirically, and what can we do about it?
- *Week 5 (Sep 28):* **Designing the validation study.** Earlier weeks examine data someone else has already collected; this week we plan our own. Power analysis, active learning, and pre-registration.
- *Week 6 (Oct 5):* **Subgroups, fairness, and privacy.** Does the deployed system work equally well for everyone? Fairness criteria, subgroup analysis, pooling when groups are small, and how privacy comes into play.