From predictions to *error-controlled* decisions

Towards actionable foundation models in medicine

Ying Jin*, Intae Moon*, and Marinka Zitnik

*Co-first authors

Ying Jin Intae Moon

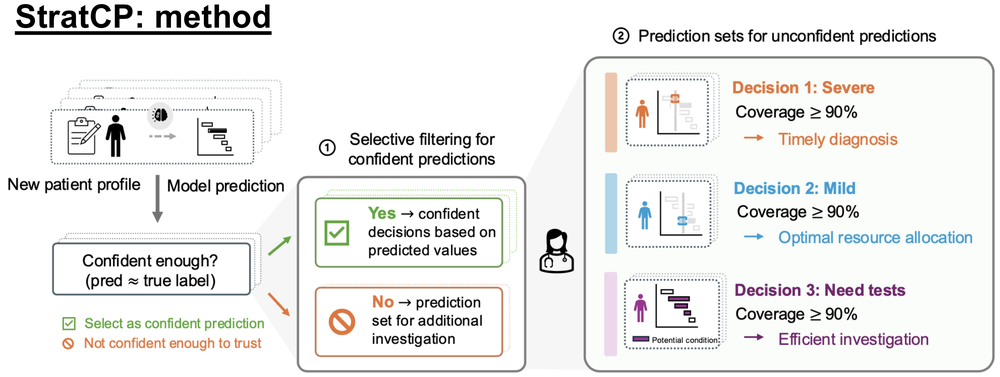
Motivation

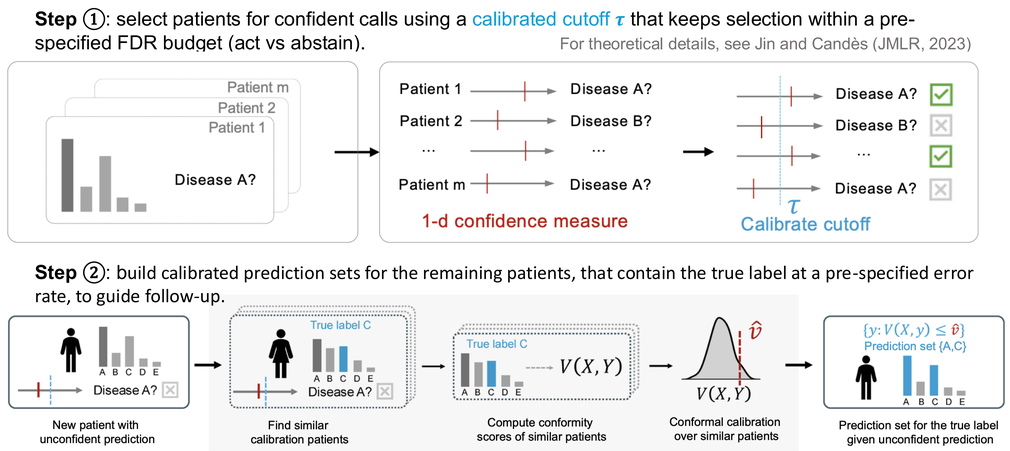
- Foundation models (FMs) show broad promise and are increasingly evaluated prospectively in medicine
- Deployment in clinical practice requires outputs that clinicians can act on under pre-specified error budgets (e.g., a cap on false-positive calls)

We introduce **StratCP**, which wraps any existing foundation model to deliver

- patient-level diagnosis/outcome within a pre-specified false discovery rate (FDR)
- calibrated, clinically coherent differentials for deferred patients

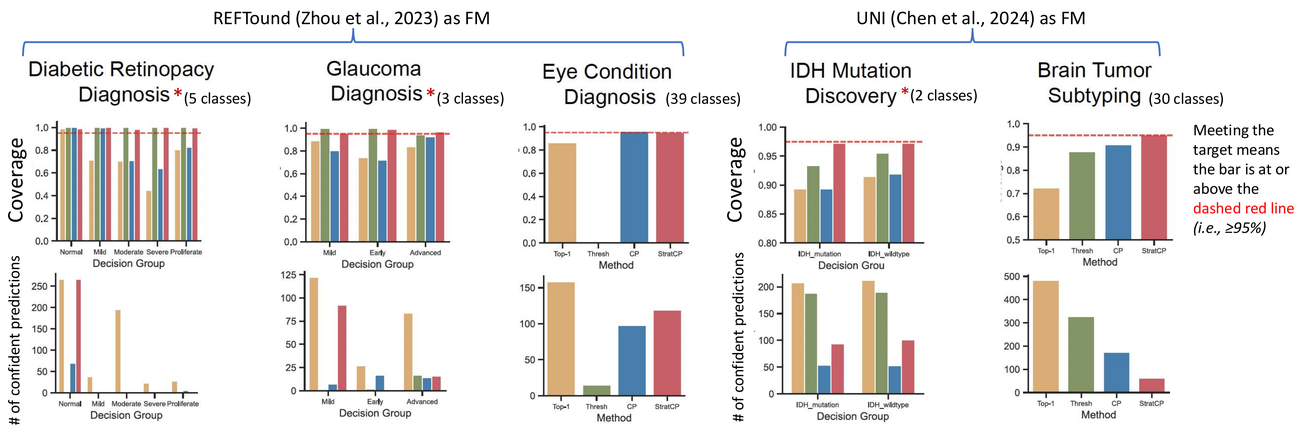
StratCP: method





StratCP: main results

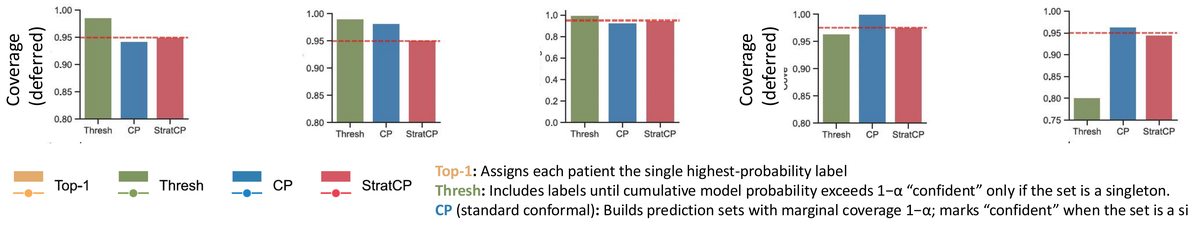
StratCP selects high-confidence patients under a pre-specified FDR budget (5% error rate) across ophthalmology and neuro-oncology, providing action-conditional guarantees*.



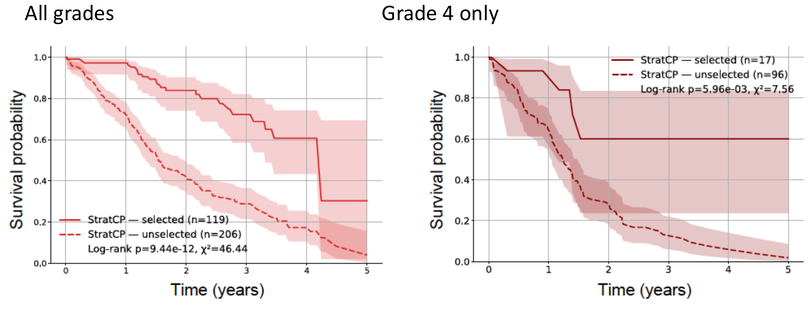
Meeting the target means the bar is at or above the dashed red line *(i.e., ≥95%)*

→ Enables the FM to *abstain* from high-confidence calls, when necessary, to meet the FDR budget.

StratCP provides calibrated prediction sets for deferred patients, which include the true label at a pre-specified error rate (5%).

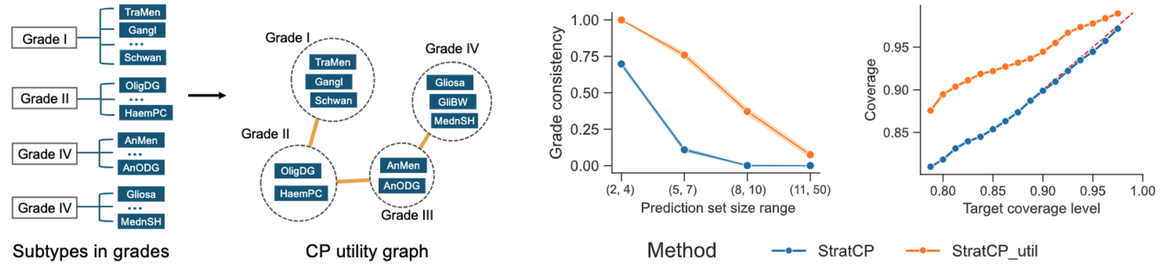


Time-to-event prediction task: StratCP's selection of long survivors (≥18 months) demonstrates clinically meaningful stratification compared to unselected (deferred) patients.

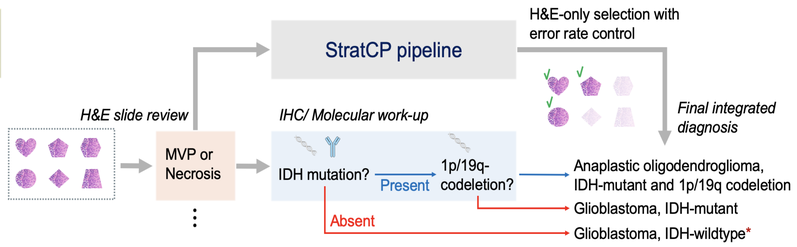


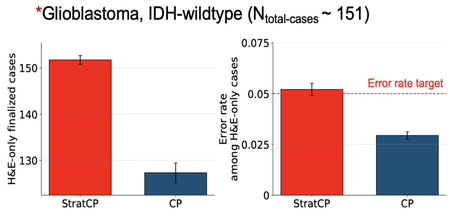
StratCP provides utility-enhanced prediction sets for deferred patients, which respect clinical adjacency and shared management, while still satisfying coverage guarantees.

Order labels using model scores + a utility graph and add the next most useful label until the calibrated coverage target is met.



For suspected adult-type diffuse glioma cases, StratCP enables H&E-only calls under a pre-specified FDR budget (5% error rate), reducing confirmatory assays and turnaround.





H&E-only finalization reduces time ⏱ and cost 💰

→ ~8.7 lab-days + $1,650 saved per case (GBM, IDH-wildtype)

→ ~66,000 lab-days + $12.5M saved annually in the US (CBTRUS 2024)

Key Takeaways:

- StratCP is a drop-in, *model-agnostic* layer that makes foundation model outputs decision-ready with per-patient error control at key points in the clinical workflow (triage, diagnosis, prognosis).
- StratCP incorporates a utility graph to generate clinically and biologically coherent prediction sets for deferred patients.
- Future work includes prospective trials (e.g., biomarker triage for IDH, TP53, and hormone receptor status) to assess StratCP in clinical deployment.