STUDENT WORKING GROUP

Newsletter

Working Group | Student

October 2026



In This Issue:

Student Spotlight:
Dr. Rachel Lee

Research Spotlight:
AgentClinic

Student Opportunities and Announcements

Find a postdoctoral fellowship

Call for Applications: AMIA STWG 2027 Leadership Roles!

1) STWG Chair-Elect (2027)
Prepare to transition into the Chair role by driving key working group initiatives, delivering high-impact symposium programming, and mentoring the next generation of informatics leaders.
**Note: Chair-Elect applicants will automatically be included on the nomination slate for the Executive Committee.

2) Executive Committee Member (2027)
Work alongside the Chair Team to guide strategy and lead projects across sub-committees, including Leadership Networking & Professional Development (Podcast Team) and Science Communication (Newsletter Team).

Deadline: Sunday, October 25th, 2026 (Anywhere On Earth)

Save The Date:

Annual Symposium

November 7-11, 2026
Dallas, TX

Amplify Informatics Conference

April 12-15, 2027
Atlanta, GA

Get connected!

AMIASTUDENTWG.SLACK.COM

@AMIASTWG

Follow the AMIA STWG on LinkedIn

STUDENT SPOTLIGHT



Rachel Yeji Lee

Assistant Professor at the Hanyang University College of Nursing

1st Place Winner of the 2026 Informatics Summit Student Paper

Please briefly describe your Student Paper Competition paper.

In our paper, “Earlier ICU Transfer after CONCERN Early Warning System Score Escalation Reduced Sepsis-related Mortality,” we conducted a post-hoc analysis of the multi-site randomized controlled trial data of the CONCERN Early Warning System to examine whether the timing of ICU transfer was associated with the risk of in-hosptial mortality among patients who later developed sepsis. We used CONCERN score escalation as a proxy for early deterioration and measured the time from escalation to unanticipated ICU transfer. Our results showed longer delays were associated with greater odds of in-hospital death, and our analysis identified 36 hours as a potential clinically actionable window for earlier escalation of care.

What was your research process and any challenges you faced in your study?

Our research process started with the question: What does an “unanticipated ICU transfer” really mean? It’s often used in the literature as an adverse outcome or a proxy for deterioration. But in our team’s recent scoping review, we found that the timing of the transfer may matter more for patient outcomes than the transfer itself. That led us to a question: then How do we define and measure an “early” versus “late” ICU transfer? One of the biggest challenges was that clinical deterioration doesn’t have a clear starting point. It’s usually a process that unfolds over time, and by the time a patient is transferred to the ICU, deterioration has often already been happening for some time.

The CONCERN RCT dataset gave us a unique way to look at that earlier trajectory because risk scores were generated hourly. We used the first score escalation from low risk to moderate or high risk as a reproducible marker of early deterioration and then measured the time from that signal to ICU transfer. That gave us a consistent starting point across patients and allowed us to examine whether the timing of ICU transfer after early signs of deterioration was associated with in-hospital mortality.

Please describe recommendations for how students can get the most out of their AMIA conference experiences.

Be open to doing more than just attending. Ask questions, talk to people, present your work, ask to be introduced, and take advantages of opportunities that come up. Conferences can certainly feel intimidating as a student, especially when it seems like everyone already knows each other or when you feel like your research is not polished enough to talk about yet. I would say you don’t wait until you feel completely “ready.” Some of the best collaborators, mentors, and friends I have in the field are people I first met through AMIA conferences, often through fairly casual conversations. I would also encourage students to take any opportunity they can to talk about their work, whether that’s through a poster, a presentation, the student paper competition, or even just a conversation with someone between sessions. For me, preparing for the student paper competition, presenting my work, and then answering questions from experts in the field was one of the most valuable parts of the conference experience. It really pushed me to think more carefully about my work and to get comfortable explaining, and sometimes defending, my ideas on the spot. One of the nice things about being a student is that you get to practice these things in a setting where you’re still expected to be learning. So take advantage of it!

RESEARCH SPOTLIGHT · npj DIGITAL MEDICINE · OPEN ACCESS

AgentClinic

Give an AI the medical exam and it scores like an expert. Make it earn the answers one question at a time—from a patient who does not volunteer them—and the picture changes.

01 PROBLEM

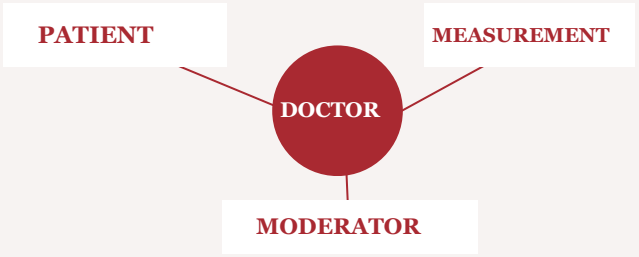
The test was too easy.

Multiple-choice exams hand the model every fact and ask it to choose. Clinics don't.

The challenge: gather the information, then make the diagnosis.

02 APPROACH

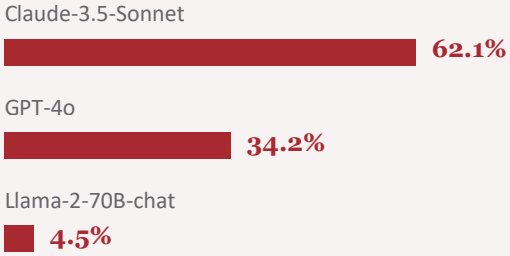
Four AI agents share one clinic.



9 specialties • 7 languages • 23 biases

03 RESULTS

Same USMLE cases, interactive format



11 models tested. Bar scale starts at zero.

04 IMPACT

Open source

MIT license—free to fork, extend and run.
More than 1,000 simulated patient cases.

Code, prompts & evaluation scripts

github.com/SamuelSchmidgall/AgentClinic

Adapted from Schmidgall et al., npj Digital Medicine (2026), licensed under CC BY 4.0.

BY THE NUMBERS

52% → 25%
Accuracy when the interaction budget was cut from 20 turns to 10

67%
Share of relevant clinical information the AI doctor actually elicited

62.1%
Best accuracy by any model on interactive USMLE cases

Doctor: Have you experienced any muscle twitching or cramps?
Patient: No, I haven't experienced any muscle twitching or cramps.

In this benchmark exchange, the patient echoes the question without volunteering new clinical details; diagnostic performance may suffer.

Full story on page 2 →

Why this matters

Medical AI is routinely judged by multiple-choice exams. Those exams hand the model every relevant fact and ask it to choose. Real clinical work is the opposite: the facts are inside the patient, and the clinician has to get them out — in limited time, with incomplete information.

AgentClinic rebuilds that reality in simulation. A doctor agent must interview a patient agent, order tests from a measurement agent, sometimes read an image, and commit to an open-ended diagnosis in 20 turns. The same questions, reformatted this way, become dramatically harder.

The reason is not that models forgot medicine. On average, doctor agents extracted only **67%** of the clinically relevant information available — and the gap between correct and incorrect diagnoses tracked that number closely (72% versus 63%). **The bottleneck is history-taking, not knowledge.**

What they did

Four AI agents share one simulated clinic. A **patient agent** knows the symptoms but not the diagnosis. A **measurement agent** returns labs, vitals and imaging on request. A **moderator agent** checks the final answer against ground truth. The **doctor agent** — the system under test — starts with almost nothing and has 20 turns to close the gap.

Cases came from USMLE questions, de-identified electronic health records (MIMIC-IV) and *New England Journal of Medicine* image challenges, spanning nine specialties and seven languages, with 23 clinical biases available to seed into either side of the conversation. Eleven models were tested as the doctor.

> *"LLMs need to be examined with novel evaluation strategies that go well beyond static question-answering benchmarks."*
>
> SCHMIDGALL ET AL., *npj DIGITAL MEDICINE* (2026)

Key takeaways

1. **Exam scores don't transfer.** Performance on the static MedQA benchmark was only weakly predictive of performance in the interactive version of the same cases.
2. **Information gathering separates right from wrong.** Doctor agents captured 67% of relevant information on average — 72% when the diagnosis was correct, 63% when it wasn't.
3. **Time is the harshest variable.** Cutting the interaction budget from 20 turns to 10 dropped accuracy from 52% to 25%.
4. **Tools help some models and hurt others.** Llama-3-70B rose from 19% to 41.1% with a persistent notebook and reflection prompting; GPT-3.5 got worse with every tool tested, losing 27.1% with textbook retrieval.
5. **Bias damages trust faster than it damages accuracy.** Implicit biases barely moved GPT-4's diagnostic accuracy but sharply reduced simulated patients' confidence, compliance and willingness to return — with education-related bias the most damaging.

What it means for you

If you evaluate or purchase clinical AI
Treat vendor benchmark scores as necessary but not sufficient. Ask for multi-turn, information-incomplete evaluations.

If you teach informatics or clinical skills
The benchmark is an OSCE analogue. It shows why history-taking, not recall, is the hard part — for humans and machines alike.

If you do research
The code, prompts and evaluation scripts are MIT-licensed and public. You can fork it and add a specialty, a language or a bias.

What this doesn't show

Every patient in the study was simulated by a language model, not a person. The human comparison rests on just three physicians, whose scores varied so widely (54% ± 28.5) that no ranking against the models is supportable. And ordering a lab test is compressed into a single command — real clinical workflows have stakeholders, dependencies and constraints that no current benchmark models.

LEARN MORE

Read the paper (open access): doi.org/10.1038/s41746-026-02674-7

Code, prompts & evaluation scripts: github.com/SamuelSchmidgall/AgentClinic

Benchmark data: open at the same repository. MIMIC-IV-derived cases require PhysioNet training and a request to the authors.