best365labs™

Your AI Pharmacist Flunked Drug Interactions — And Nobody Noticed.

Picture this: A patient takes a 10mg methylene blue supplement every morning. She's also on Lexapro for anxiety. Smart woman. Health-conscious. So before her next appointment, she does what millions of people do now — she asks ChatGPT.

The AI responds with something like: "⚠️ DANGER: Combining methylene blue with SSRIs like escitalopram (Lexapro) carries a serious risk of serotonin syndrome, a potentially life-threatening condition. This combination should be avoided."

She panics. Stops her methylene blue cold. Calls my office in tears. Comes in convinced she's been poisoned.

Here's the thing: ChatGPT wasn't exactly right. It wasn't exactly wrong either. It was something worse — it was confidently, dangerously oversimplified. And that's the AI drug interaction problem in a nutshell.

The Numbers Are Damning

Let's not bury the lede. AI chatbots are genuinely terrible at drug-drug interactions (DDIs). Not slightly off. Not "needs improvement." Terrible.

A 2024 study in the British Journal of Clinical Pharmacology put ChatGPT-3.5 through its paces using real hospitalized patient data. The result? It missed roughly **75% of actual drug-drug interactions**. Cohen's kappa agreement between ChatGPT and trained pharmacists ranged from just 0.077 to 0.143. For reference, kappa values below 0.2 are considered

“slight agreement” — barely better than random chance. [1]

A coin flip. For something that can kill people.

A 2025 study in Exploratory Research in Clinical and Social Pharmacy tested ChatGPT, Gemini, and Microsoft Copilot against 57 real patient medication lists encompassing 204 known DDIs. ChatGPT’s precision was **19%**. Let that land for a second. That means 4 out of every 5 drug interaction warnings ChatGPT generated were false positives. The best F1 score across all platforms was 0.25. The authors’ conclusion was unambiguous: *“No AI systems assessed achieve the required balance of precision and sensitivity for reliable clinical decision-making.”* [2]

And it's not just interactions. A 2023 study found ChatGPT accurately answered only about **25% of medication questions overall**. It fabricated a dose conversion factor out of thin air. It cited medical organizations that never published the guidance it attributed to them. [3]

The American Society of Health-System Pharmacists found that pharmacists reviewing ChatGPT’s drug-related answers flagged nearly **75% as incomplete or inaccurate**. In one case, ChatGPT told a user that Paxlovid and verapamil had no clinically significant interaction. They absolutely do — the combination can cause a dangerous drop in blood pressure. That’s not a minor error. That’s a hospitalization waiting to happen. [4]

A 2025 study in the *British Journal of Clinical Pharmacology* comparing physician responses to ChatGPT on real pharmacotherapy queries found ChatGPT “substantially inferior.” Among the gems: ChatGPT described tazobactam — a beta-lactamase inhibitor — as a “sedative agent.” It described Actrapid, which is **insulin**, as an “analgesic.” These aren't rounding errors. These are category-level confabulations.[5]

The Methylene Blue Problem: A Perfect Case Study

Back to my patient.

Methylene blue has a real pharmacological interaction with serotonergic drugs — I'm not here to pretend otherwise. It inhibits MAO-A, which breaks down serotonin, so in theory, combine it with an SSRI and you could push serotonin levels higher. The concern is legitimate.

But the context is everything. And AI doesn't do context.

The FDA’s own 2011 safety communication — which is the foundation for nearly every chatbot warning on this topic — was triggered by cases involving **intravenous methylene blue** given during surgical procedures, specifically parathyroid surgeries, at doses ranging from 1 to 8 mg/kg IV. The FDA’s explicit language about oral dosing? *“It is not known whether there is a risk of serotonin syndrome in patients taking serotonergic psychiatric medications who are given methylene blue by other routes (e.g., orally or by local tissue injection).”* [6]

Not known. The FDA said not known.

The oral supplement world operates at 0.5 to 2 mg/kg — doses far below the IV surgical range, through a completely different pharmacokinetic pathway. IV administration delivers rapid, high plasma peaks. Oral administration is absorbed gradually through the GI tract. Same molecule. Completely different exposure profile.

How many case reports exist of oral methylene blue causing serotonin syndrome? According to a comprehensive review by Fagron Academy, exactly **one**. A single case report. Meanwhile, thousands of patients have been taking oral compounded methylene blue with no formally documented cases of serotonin syndrome in that population. [7]

What does AI do with this nuance? It ignores it entirely. It sees “methylene blue” + “SSRI” and outputs a maximum-severity red flag. Route of administration? Irrelevant. Dose? Doesn't matter. The difference between 1 mg/kg IV and 10mg oral? Flattened into a single warning emoji and a liability disclaimer.

The Double Failure Nobody Talks About

Here's what makes AI DDI performance uniquely dangerous: it fails in both directions simultaneously.

It **misses** real interactions — like the Paxlovid/verapamil case where it said no problem. And it **overcalls** non-interactions — blasting patients with terrifying warnings about theoretical risks based on mechanism extrapolation rather than clinical evidence.

Miss rate of ~75% on real interactions. False positive rate of ~80% on flagged interactions. Both in the same tool. At the same time.

A 2025 study from Mount Sinai captured the other half of this problem: AI chatbots are “highly vulnerable to repeating and elaborating on false medical information.” They don't just parrot wrong information. They build on it. A single made-up term can trigger a confident, detailed, internally coherent explanation about a completely fictional condition. Researchers found a single one-line warning prompt cut hallucinations by nearly half. One line. Which means without that line, the default mode is confident confabulation.[8]

Why This Keeps Happening

This isn't a mystery. It's a design problem.

Large language models are trained to generate responses that are coherent, authoritative-sounding, and — critically — defensible. When in doubt about a drug interaction, the safest output for the model is to warn. Always warn. Maximum caution. If it warns and the interaction is real, it looks responsible. If it warns and the interaction is theoretical, it looks appropriately conservative. If it fails to warn and something goes wrong, it looks catastrophic.

So AI optimizes for liability, not accuracy.

Real clinical pharmacology doesn't work that way. Real pharmacology demands dose-response relationships, route of administration, patient-specific factors, mechanism of action, and actual case report evidence. It demands nuance. It demands saying “the risk here is theoretical and the evidence base is a single case report” rather than “DANGER: avoid combination.”

And here's the kicker: even the best-performing model, ChatGPT-4o, was judged to “not ensure sufficient accuracy and completeness of responses, which significantly limits its practical application.” That's the best of the bunch. Still not good enough for clinical use. [9]

What Physicians Actually Need to Do

First: Talk to your patients about this. They are using these tools. They are making decisions based on them. When patients bring AI-generated concerns into the exam room — and they will — be ready to engage with the evidence, not just dismiss the question.

Second: Consider the evidence hierarchy when AI and reality conflict. One case report of oral methylene blue causing serotonin syndrome — ever — is not the same as a black box warning. Route, dose, and mechanism matter. Teach patients to ask: “What dose? What route? How many actual cases?”

Third: If you're building clinical decision support tools, stop treating off-the-shelf chatbots as substitutes. Custom GPTs with curated, evidence-grounded pharmacology databases are meaningfully different from general-purpose AI. The technology isn't the problem — the deployment is.

Fourth: Publish and advocate. The evidence that AI DDI screening is dangerously unreliable is now substantial. The AMA and FDA regulatory frameworks around AI in clinical decision-making are still catching up to reality. Push for standards. Support outcome-based benchmarks for AI clinical tools. “It sounds confident” is not a clinical standard.

The Bottom Line

Your patients are trusting AI chatbots with medication safety decisions. Those chatbots miss three out of four real drug interactions. They generate false alarms four times out of five when they do flag something. They call insulin an analgesic. They elaborate confidently on things they've fabricated. And the best one currently available still doesn't meet the bar for clinical reliability.

Meanwhile, my patient stopped a supplement with genuine cognitive and mitochondrial benefits because an AI didn't know the difference between an IV surgical dose and a 10mg tablet.

We built these tools. We deployed them. We haven't told patients they're not ready. That's on us.

The AI drug interaction problem isn't coming. It's here. And the first step to fixing it is admitting it exists.

Dr. Steve Warren is a quad-board certified physician in Family Medicine, Preventive Medicine, Addiction Medicine, and Hospice and Palliative Medicine, with a clinical focus on longevity medicine. He is the founder of Best 365 Labs, which develops evidence-informed longevity supplements including oral methylene blue. He writes at the intersection of emerging therapeutics and clinical evidence, calling out hype where he sees it — and sounding the alarm when the hype machine runs in the other direction. He publishes regularly on KevinMD.

Citations:
1. British Journal of Clinical Pharmacology (2024), https://pmc.ncbi.nlm.nih.gov/articles/PMC11602951/
2. Exploratory Research in Clinical and Social Pharmacy (2025), https://pmc.ncbi.nlm.nih.gov/articles/PMC12712589/
3. CNN Health (2023), https://www.cnn.com/2023/12/10/health/chatgpt-medical-questions
4. Fox Business / ASHP Study (2023), https://www.foxbusiness.com/technology/study-finds-chatgpt-provided-inaccurate-answers-medication-questions
5. British Journal of Clinical Pharmacology (2025), https://pmc.ncbi.nlm.nih.gov/articles/PMC12930016/
6. FDA Drug Safety Communication (2011), https://www.fda.gov/drugs/drug-safety-and-availability/fda-drug-safety-communication-updated-information-about-drug-interaction-between-methylene-blue
7. Fagron Academy (2025), https://www.fagronacademy.us/blog/methylene-blue-understanding-drug-interactions
8. Mount Sinai Health System (2025), https://www.mountsinai.org/about/newsroom/2025/ai-chatbots-can-run-with-medical-misinformation-study-finds-highlighting-the-need-for-stronger-safeguards
9. Journal of Clinical Medicine (2025), https://pmc.ncbi.nlm.nih.gov/articles/PMC12608490/