



Physician Ratings of Clinical AI Tools

Executive Summary

AI-based clinical reference tools are increasingly used at the point of care. This report summarizes results from 1,315 side-by-side evaluations in which physicians compared responses from DoxGPT with those from other clinical AI tools (e.g., OpenEvidence, UpToDate, ChatGPT) using clinical questions they personally posed on each platform.

Across evaluations, most questions produced clinically similar answers—67% marked the preferred clinical answer as “Nearly the same” or “Bit better”. This is expected given their reliance on comparable foundational language models.

When a preference was expressed, DoxGPT was favored at more than twice the rate of the nearest competitor (61% vs 26%) when tallying which answer was best. Reasons cited for DoxGPT included:

- Full PDF access to over 2,000 medical journals
- Built-in deterministic drug reference
- Absence of full-screen banner ads
- Tables and formatting that were easier to read
- Faster response times

DoxGPT was most often preferred for drug-related questions and for complex or niche evidence queries, reflecting the value of deterministic drug references and direct access to full-text primary literature.



Overall Ratings

DoxGPT was selected as the best clinical answer in 61% of side-by-side evaluations.

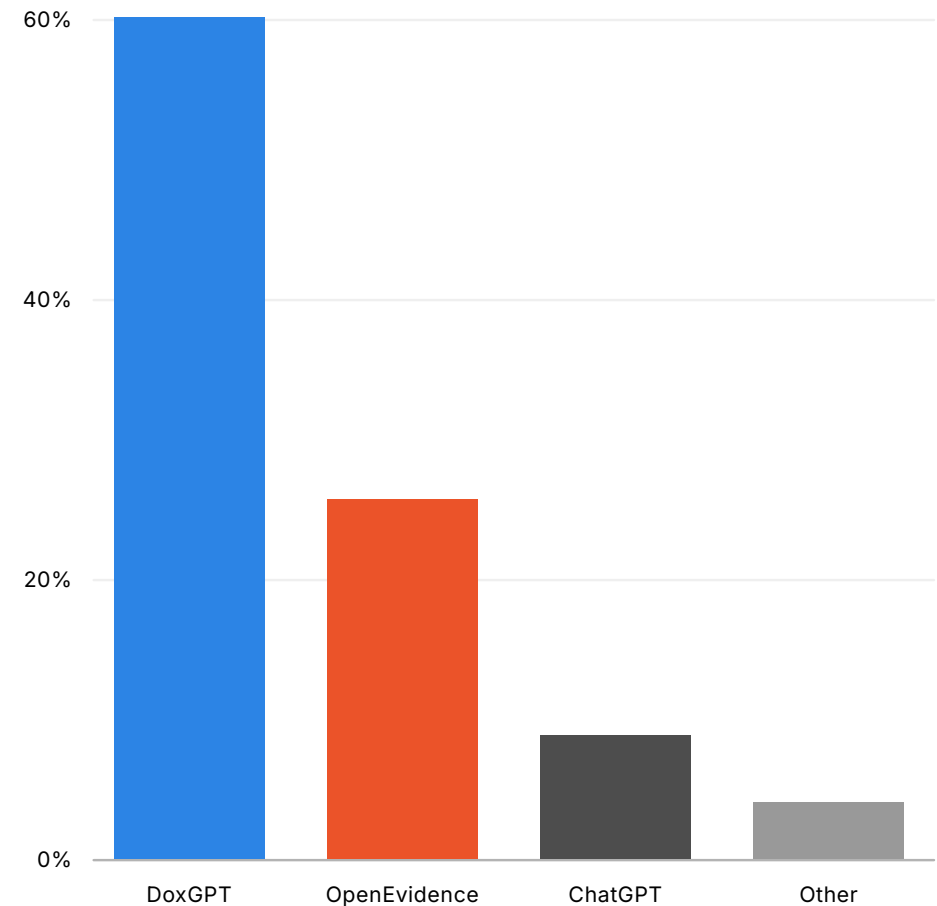
OpenEvidence accounted for most alternative selections (26%), followed by ChatGPT (10%).

61%

DoxGPT was selected as
the best clinical answer

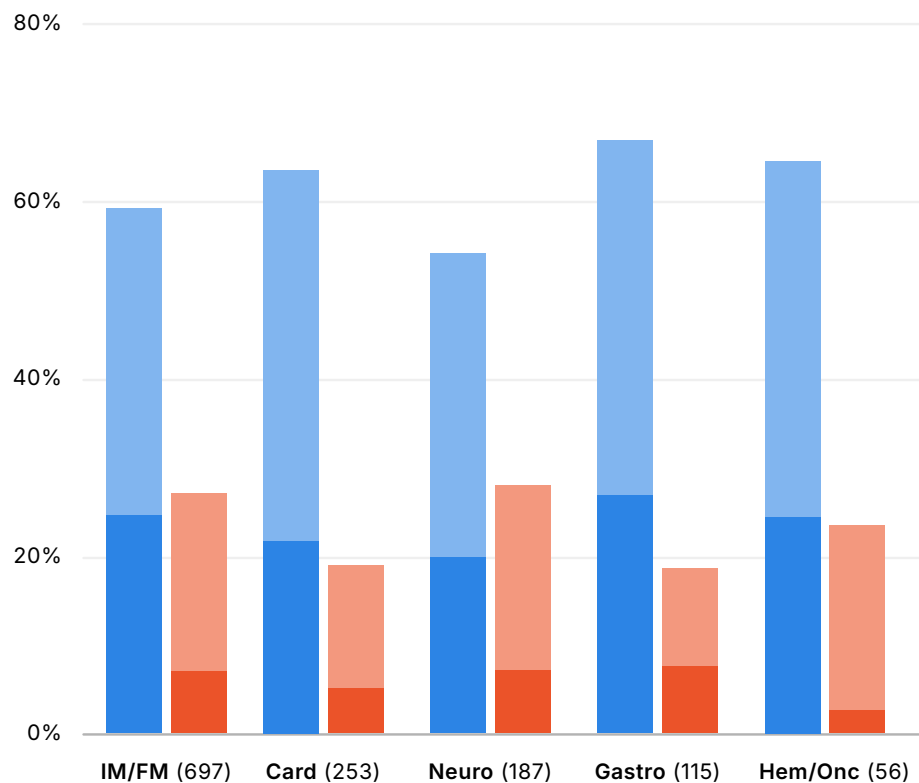
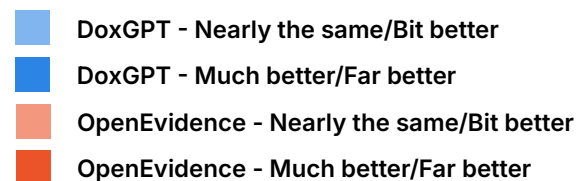
Which AI answered your clinical question the best?

(By Tool) | n = 1,315



Which AI answered your clinical question the best?

(By Specialty) | n = 1,315



Specialty Ratings

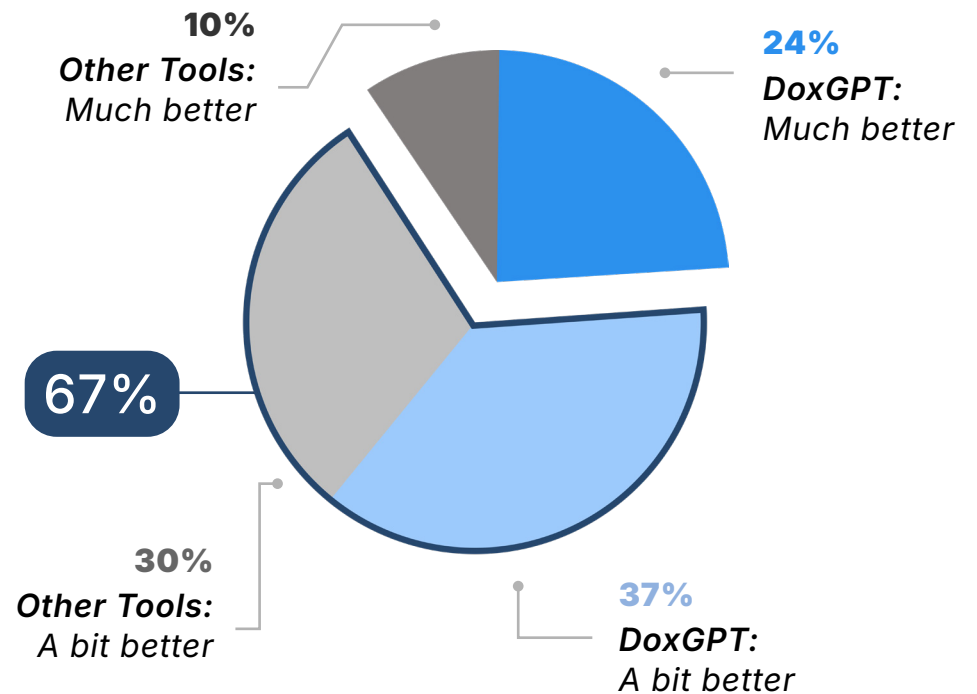
DoxGPT performed best with specialists where access to thousands of niche journals helps most.

- Gastroenterology (67%)
- Hematology/Oncology (66%)
- Cardiology (64%)
- Family Medicine (59%)
- Neurology (56%)

Similar Sources & Models

Most questions (67%) yielded answers that were “Nearly the same” or a “Bit better” across tools, a product of high correlation among sources and foundational models used.

Magnitude of difference across tools*



*"Bit Better" represents respondents who selected either "Bit better" or "Nearly the same", and "Much better" represents respondents who selected "Much better" or "Far better".

Methods

A total of 1,315 independent side-by-side evaluations were completed between November 17, 2025 and January 22, 2026. During these evaluations, physician specialists compared DoxGPT with an alternative clinical AI tool of their choosing for clinical reference questions they posed based on their practice.

All respondents were first-time DoxGPT users, and were not incentivized to answer in any particular way.

For each comparison, respondents selected:

1. The tool providing the better clinical answer
2. The magnitude of difference (*"Nearly the same"*, *"Bit better"*, *"Much better"*, *"Far better"*)
3. A free-text explanation of their choice

Our goal is more transparency in clinical AI. If interested in specific comparative examples or broader access to this dataset, please inquire at: pr@doximity.com.



Limitations

Results are based on self-reported user assessments and may be influenced by familiarity or interface differences. Qualitative themes reflect explicit mentions and may underrepresent implicit judgments.

Participants received compensation for completing evaluations, which may influence response behavior. In addition, all respondents were members of the Doximity platform. However, Doximity membership includes over 80% of U.S. clinicians, making this sample broadly representative of the practicing clinician population likely to encounter AI-based clinical reference tools.