

Introduction

Vestibular schwannomas (VS) are benign tumors arising from Schwann cells of the vestibulocochlear nerve (CN VIII), accounting for 8% of intracranial tumors. Treatment paradigms have evolved with advances in imaging and surgical technology, and treatment frequency has increased due to expanded MRI screening for asymmetric hearing loss.

- Clinical guidelines for VS treatment are NOT standardized, which can lead to variability in outcomes. Carlson et al. 2020 developed 103 expert consensus statements via modified-Delphi methodology.

- Artificial intelligence (AI) may help healthcare providers collate expert opinions and current literature for clinical decisions, but thorough evaluation is necessary before implementation.

Platforms vary in training data, algorithms, and citation accuracy, which may impact reliability for medical applications.

- Objective:** This study evaluates whether widely available AI tools agree with expert consensus on VS management, assessing performance across multiple platforms in 2023 and 2025 to determine feasibility for clinical decision-making support.

Methods

Original Platforms (Initial Assessment):

- scite.ai (Scite, Brooklyn, New York, USA)
- ChatGPT4 (OpenAI, San Francisco, California, USA)
- Bard (Google, Mountain View, California, USA)

2025 Successor Platforms

- SCITE - updated version of scite.ai
- GPT-4o - successor to ChatGPT4
- Google Gemini - successor to Bard
- DeepSeek

Platform Selection Criteria

Widely recognized AI assistants, both free and paid subscription tiers and functionally independent from each other.

Standard Expert Consensus Comparison

- Carlson et al. 2020 modified-Delphi consensus study
- Publication date allows assessment of AI ability to process older data relative to training cutoffs

Study Design

- Each consensus statement converted to question format (see Table 1)
- Each AI queried independently with identical questions
- AI responses classified as: Agree / Disagree / Uncertain with 2 person consensus
- All queries performed via browser interface

Outcome Measures

- Primary: Agreement rate (%) with expert consensus per AI platform
- Secondary: Agreement rate by consensus category
- Comparison of performance across AI platforms

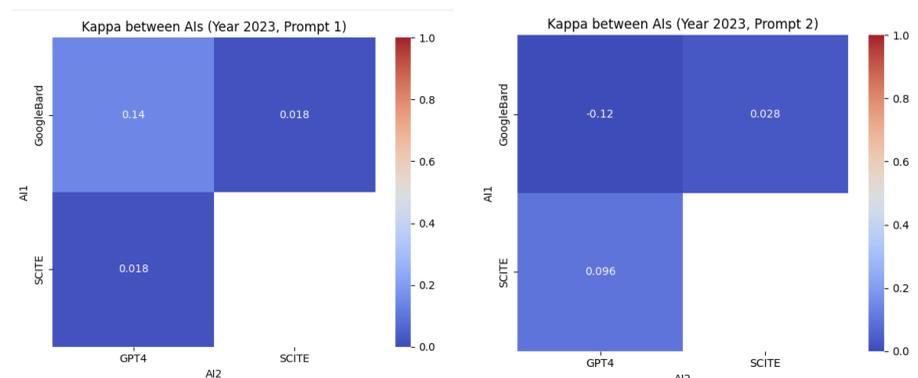
Results

Table 1. Example of Responses to Prompts by Different AI Models

Platform	Example Statement	Prompt 1: Do you agree with the following statement?	Prompt 2: Rephrase as a Yes/No question and answer.
Bard	The overall probability of maintaining serviceable hearing ($\leq 50\%$ WRS) following single-fraction radiosurgery for a vestibular schwannoma that is less than 2 cm in greatest posterior fossa diameter, utilizing modern dose planning (12–13 Gy marginal dose), is 60–80% at 2 years, 40–60% at 5 years, and 10–30% at 10 years following treatment.	Yes, I agree with the statement that the overall probability of maintaining serviceable hearing following single-fraction radiosurgery for a vestibular schwannoma is as stated.	Rephrased question: "Is it likely that a patient with a vestibular schwannoma that is less than 2 cm in greatest posterior fossa diameter will have serviceable hearing 10 years after single-fraction radiosurgery with modern dose planning?" Answer: No, the probability at 10 years is only 10–30%, making hearing loss more likely.
Scite	(Same as above)	No, I do not agree with the statement. The provided probabilities are not supported by the available evidence.	Rephrased question: "Can the probability of maintaining serviceable hearing ... following single-fraction radiosurgery ... be summarized as 60–80% at 2 years, 40–60% at 5 years, and 10–30% at 10 years?" Answer: Yes, according to Pollock et al., the probability can be estimated as 70–90% at 2 years, 40–60% at 5 years, and 20–40% at 10 years.
GPT-4	(Same as above)	There is no comprehensive long-term consensus on these exact probabilities. The figures may be accurate but might not reflect the full range of possible outcomes.	Rephrased question: "Is it accurate to say that the probability of maintaining serviceable hearing ... following single-fraction radiosurgery ... is 60–80% at 2 years, 40–60% at 5 years, and 10–30% at 10 years?" Answer: Yes, it is generally possible for hearing preservation rates to fall within the mentioned ranges with modern dose planning.

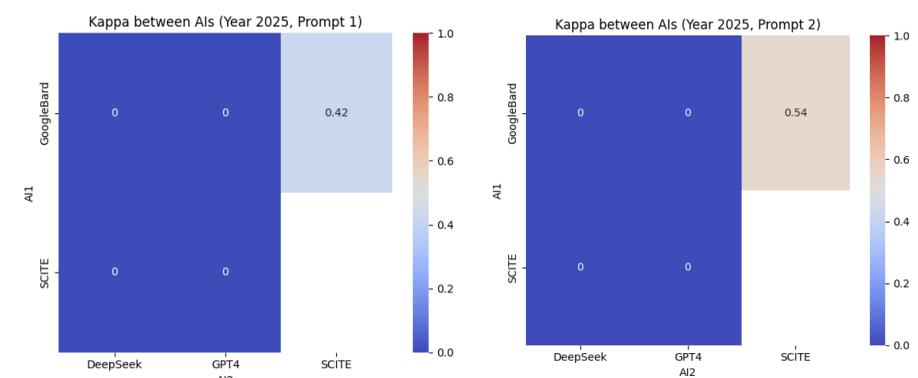
2023 Expert Consensus Averaged per Statement Category:

- Google Bard: 84.5%
- GPT-4: 88.8%
- SCITEAI: 57.3%
- Range: 45.6% to 92.2%



2025 Expert Consensus Averaged per Statement Category:

- Google Gemini: 93.2%
- GPT-4o: 100%
- SCITE 2025: 91.7%
- DeepSeek: 100%
- Overall range: 90.3% to 100%



Results (cont.)

1. Within-AI Consistency (Cohen's Kappa)-----

- 2023: Kappa ranged 0.039 to 0.371 (slight to fair agreement)
- 2025: Kappa ranged 0.297 to 1.000 (increased internal consistency)

2. Between-AI Agreement (Cohen's Kappa)-----

2023 Inter-AI Kappa (poor agreement):

- Bard vs. SCITEAI: 0.018 (P1), 0.028 (P2)
- Bard vs. GPT-4: 0.142 (P1), -0.118 (P2)
- SCITEAI vs. GPT-4: 0.018 (P1), 0.096 (P2)

2025 Inter-AI Kappa (improved):

- Bard vs. SCITE: 0.425 (P1), 0.540 (P2) - moderate agreement
- GPT-4 vs. DeepSeek: Not Applicable (identical responses - all correct)

Conclusions

As expected, AI performance improved substantially from 2023 to 2025. Easily accessible AI tools can provide answers that vary significantly depending on the AI algorithm and specific prompts used, but GPT-4o and DeepSeek achieved perfect (100%) agreement with expert consensus on VS management.