

Whisper

The use of generative artificial intelligence-based dictation in a neurosurgical practice: a pilot study

In a [pilot comparative study](#) Hopkins et al. from the Keck School of Medicine, USC, Los Angeles (Neurosurgery; Endocrinology) published in [Neurosurgical Focus](#) to evaluate whether a modified [OpenAI Generative artificial intelligence](#) model can match or improve upon the accuracy of a commercial dictation tool (Nuance Dragon Medical One) in [neurosurgical operative report](#) generation. [Whisper](#)-based model demonstrated non-inferior overall word error rate (WER) versus Dragon (1.75% vs 1.54%, $p=0.08$). Excluding linguistic errors, Whisper outperformed Dragon (0.50% vs 1.34%, $p<0.001$; total errors 6.1 vs 9.7, $p=0.002$). Whisper's performance was robust to faster speech and longer recordings, unlike Dragon ¹⁾.

Critical Review

* Strengths:

- Direct comparison of a cutting-edge generative AI (Whisper) to an established clinical tool in a real-world neurosurgical workflow.
- Objective metrics (WER) with appropriate statistical analysis.
- Mixed-case operative reports cover cranial and spinal procedures, enhancing generalizability.

* Weaknesses & Limitations:

- Small sample size ($n=10$ reports, 3 physicians) limits statistical power.
- Lack of real-time clinical integration assessments—only offline comparisons.
- No analysis of downstream impact on report quality, clinician satisfaction, or patient safety.
- Commercial Dragon may not represent the latest version or fully optimized settings.

* Methodological concerns:

- Manual WER calculation introduces potential reviewer bias; no inter-rater reliability reported.
- Recording conditions and audio quality not standardized across sessions.
- Exclusion of “linguistic errors” may bias interpretation toward AI advantage.

* Clinical relevance:

- Whisper's stability with faster dictation may support efficiency gains in high-volume clinical settings.
- Noninferiority demonstrated, but real-world deployment needs integration, EHR compatibility, user training, and error recovery workflow.

Verdict: 6.5 / 10

Criteria	Score	Comments
Innovation	8	Novel application of transformer-based AI to dictation
Methodology	6	Solid but limited by small sample and manual error assessment
Clinical Applicability	6	Promising, yet lacks prospective implementation data
Statistical Rigor	6	Basic significance testing performed; confidence intervals absent

Key Takeaway for Neurosurgeons: Modified [Whisper](#) offers comparable, or potentially better, transcription accuracy than Dragon in neurosurgical dictation, especially under faster speech rates—but further large-scale, workflow-integrated trials are essential before clinical adoption.

Bottom Line: This pilot suggests generative AI could reduce documentation burden, but robustness and clinical utility must be validated in real-world settings.

Title: The use of generative artificial intelligence-based dictation in a neurosurgical practice: a pilot study **Full Citation:** Hopkins BS, Dallas J, Yu J, Briggs RG, Chung LK, Cote DJ, Gomez D, Shah I, Carmichael JD, Liu JC, Mack WJ, Zada G. **Neurosurg Focus**. 2025 Jul 1;59(1):E8. doi:10.3171/2025.4.FOCUS24834 **Publication Date:** July 1, 2025

Categories: Research, AI in Neurosurgery, Clinical Workflow, Pilot Studies **Tags:** generative AI, transcription, Whisper model, Dragon Medical One, neurosurgery documentation, word error rate

¹⁾
Hopkins BS, Dallas J, Yu J, Briggs RG, Chung LK, Cote DJ, Gomez D, Shah I, Carmichael JD, Liu JC, Mack WJ, Zada G. The use of [generative artificial intelligence](#)-based [dictation](#) in a [neurosurgical practice](#): a [pilot study](#). *Neurosurg Focus*. 2025 Jul 1;59(1):E8. doi: 10.3171/2025.4.FOCUS24834. PMID: 40591970.

From:

<https://neurosurgerywiki.com/wiki/> - **Neurosurgery Wiki**

Permanent link:

<https://neurosurgerywiki.com/wiki/doku.php?id=whisper>

Last update:

2025/07/06 20:53

