# Comparison of Provider-Generated vs Artificial Intelligence-Generated Medical Encounter Notes

Edward Harrington[1], Francisco Cordero[1], Michelle Cangiano, MD[1,2,3], Alicia Jacobs, MD[1,2,3]
The Robert D Larner MD College of Medicine at the University of Vermont[1] Department of Family Medicine[2], UVMHN[3]

## Background

UVMHN AI Pilot Study

This Spring UVM Health Network carried out a double crossover Pilot Study wherein 50 primary care providers utilized 2 different Artificial Intelligence (AI) generating note programs to evaluate impact on provider wellbeing.

Over a 2-month period, both vendors improved provider fulfillment at work and decreased burn-out.

Note Quality

Previous studies have use a modified PDQI-9 to include AI-specific traits such as hallucination and bias, and a 5-point Likert scale[1].

| Note Quality Categories |
|---|
| Accurate and correct |
| Thorough |
| Useful |
| Organized and complete |
| Comprehensible |
| Succinct and consider |
| Synthesized with contingency and clinical assessment |
| Internally consistent |
| Free from hallucinations |
| Free from bias |
| Readability |

## Hypothesis

This study aims to determine the quality of AI-generated notes in comparison to Provider-generated notes within UVMHN. We also want to determine if there is difference in note quality between the two AI scribe programs.

Primary

There is no difference in quality between AI-generated and Provider-generated medical encounter notes.

Secondary

There is no difference between in quality between the Vendor "X" and Abridge AI-generated notes.

## Study Design

- Identified providers enrolled in AI pilot study
- Located and classified all out-patient, AI-generated notes in Epic.
- A power calculation estimated 46 notes needs for our primary analysis and 23 notes needed for our secondary analysis
- We excluded the following types of notes: PE, wellness, pre-ops, video, well-child checks.
- Randomly (random number generator) selected “y” such that an equal percentage of notes from each classification category were chosen, and the total sum of notes equals the total required to reach desired power. The purpose of this was two-fold: first we wanted to ensure our sample set was representative of our total note population; second this allows for us to engage in follow-up analysis on these sub-categories.
- For each AI note included in our analysis we used Epic to identify the most recent (but prior to the start of the pilot study) provider-generated note produced by that AI note’s provider, that matched the relevant classification categories outlined above
- For note-rating, both raters used our rating tool on a training set to identify differences in our rating technique.
- Each rater evaluated all of the notes in our study set.
- We analyzed the data to determine how AI-generated notes compared to provider-generated notes overall and within the classification categories outlined above.

## Results

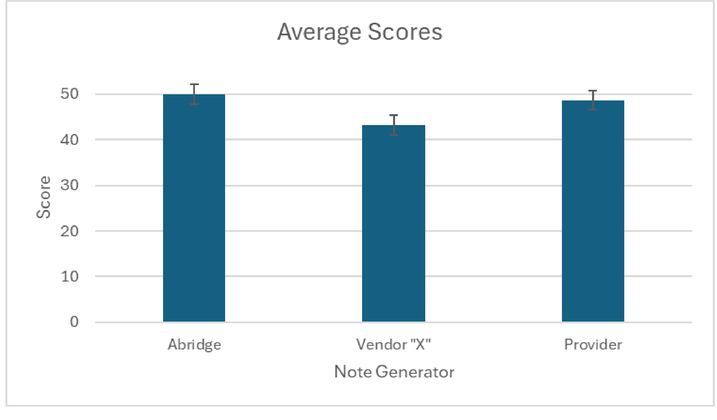
The means of the note scores from the provider-generated notes were compared to those of the AI-generated notes using a paired t-test.

Note Quality Results:

| Training Set Results | Average Score | Standard Deviation |
|---|---|---|
| Abridge-generated | 49.0 | 2.21 |
| Provider-generated | 49.4 | 3.53 |

| Note Generator | Average Score | Standard Deviation |
|---|---|---|
| Abridge | 50.0 | 2.16 |
| Vendor "X" | 43.2 | 2.20 |
| Provider | 48.6 | 2.06 |

| Comparison | t-value | p-value |
|---|---|---|
| Abridge vs. Provider | 1.48 | 0.16 |
| Abridge vs. Vendor "X" | 4.92 | 0.0008 |




No difference in quality between Abridge-generated notes versus Provider-generated notes was detected.

A difference in quality between Abridge-generated notes and Vendor "X"-generated notes was detected; specifically, that Abridge-generated note were of better quality than Vendor "X"-generated notes.

## Discussion/Conclusion

Abridge AI generated notes were non-inferior to provider generated notes.

Not all AI vendor notes are the same quality.

Using generative AI as a wellbeing intervention does not impact quality of notes.

Future work includes identifying if this quality extends to all note types and all specialty documentation.

Utilizing AI to rate medical notes was out of scope for this study but may be the subject of future studies within UVMHN.

Acknowledgements: Drs. Yao Li, Marie Sandoval, Sean Maloney, Rachel K McEntee

## References

1.Tierney, A, A. Ambient Artificial Intelligence Scribes to Alleviate the Burden of Clinical Documentation. NEJM, Vol.5 No. 3, March 2024. DOI: 10.1056/CAT.23.0404
2.Podder, V., et al. SOAP Notes. NCBI Bookshelf, StatPearls, Jan. 2024.
3. Stetson, P.D., et al. Assessing Electronic Note Quality Using the Physician Documentation Quality Instrument (PDQI-9). Applied Clinical Informatics. 2012, 3: 164 – 174. Doi:10.4338/ACI-2011-11-RA-0070.
4.Feldman, J., et al. Scaling Note Quality Assessment Across an Academic Medical Center with AI and GPT-4. NEJM Catalyst, Innovations in Care Delivery. Vol. 5, No. 5, May 2024.
5. Huang, J., et al. An Empirical Evaluation of Depp Learning for ICD-9 Code Assignment using MIMIC-III Clinical Notes. Computer Methods and Programs in Biomedicine. June 11, 2019.
6. Vasey, B., et al. Reporting guideline for the early stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. BMJ, Research Methods and Reporting. 2022; 377:e070904, doi: 10.1136/bmj-2022-070904.
7.Han, R., et al. Randomized controlled trials evaluating artificial intelligence in clinical practice: a scoping review. The Lancet. Vol. 6, May 2024.

*Supported by HRSA U77HP03624 and the VT AHEC Scholars Program, focus area Medical Practice Transformation*