Leveraging Automatic Speech Recognition (ASR) and hybridized transcription to improve scale and processing time while maintaining human-level accuracy for speech biomarkers

Rachel Kindellan, Rachel Newsome, Jordan Ponn, Aaron Rambhajan, Sasha Sirotkin, Bill Simpson, Celia Fidalgo

CAMBRIDGE COGNITION

WINTERLIGHT

Objective

Leveraging speech and language digital biomarkers in clinical development depends on accurate transcription. The gold standard approach using highly trained transcriptionists is accurate, but time consuming and challenging to scale. Automated approaches using automatic speech recognition (ASR) algorithms are significantly faster and highly scalable, but are lower accuracy. Our goal was to investigate the speed and accuracy of manual, semi-automated, and fully automated transcription of different types of clinical speech samples.

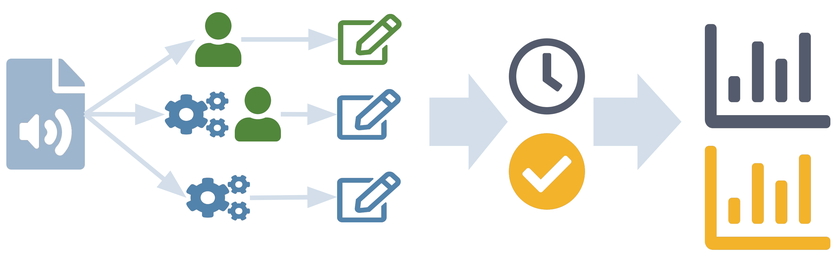
Design

We leveraged a dataset of 8 audio samples containing either open-ended speech (N=4) or verbal fluency tasks (N=4). Each sample was transcribed three ways: (1) manually by a team of human transcriptionists, (2) semi-automatically with human-review of ASR, and (3) fully automatically with ASR only. We analyzed the speed (ratio of work duration to raw audio duration) and accuracy (word error rate; WER) compared to a gold standard transcript. We then leveraged a second dataset (N=1669) of both open-ended and fluency tasks to further validate potential differences in processing time between manual and semi-automatic approaches.

A

	Fluency	Open-ended
Instruction	"*List* as many words as you can starting with the letter s"	"*Describe* this picture"
Response (transcribed)	"sun salt sand sea sieve silt"	"I see a dog with a spatula in its mouth"

B



C

Method	Words	Hesitations	Annotations	QA flags
Manual	✓	✓	✓	✓
Semi-auto	✓	✓	✓	✓
ASR	✓	✓	✗	✗

Figure 1A: Examples of transcribed speech in fluency tasks and open-ended tasks. Figure 1B: Methodology Schematic. Each audio sample was transcribed three ways: manually, semi-automatically, and automatically. Transcription speed and transcript accuracy were then analysed. Figure 1C: Table of transcript features available for each transcription method. Annotations capture stutters, word fragments, non-words, and other complex linguistic and non-linguistic features. Quality Assurance (QA) markers describe the integrity of the audio file and recording environment as well as capture variation in rater behaviour when applied to recording of clinical interviews.

Results

Overall, manual and semi-automatic approaches were significantly more accurate than ASR (p=0.02). However, this difference depended on the type of speech sample, with ASR achieving comparable accuracy to manual for open ended, but not fluency tasks (Figure 2A). Semi-automatic transcription was also significantly faster than manual overall (Figure 2B, p=0.005), but similar to accuracy, we observed an interaction whereby semi-automatic transcription yielded significant speed gains for open-ended speech samples while fluency tasks were slower. This was due to the higher volume of errors (lower accuracy) produced by the first pass ASR transcript, resulting in more time required to make corrections.

To understand the best fit for fully manual transcription within a speech processing pipeline, we further evaluated speed differences between manual and semi-automatic approaches in a much larger validation dataset (n=1669) of open-ended and fluency tasks (Figure 2C). Consistent with the pilot dataset, we observed a significant, 10.4% speed advantage over manual for open-ended tasks (p <0.001), but not for fluency tasks (p = 0.453).

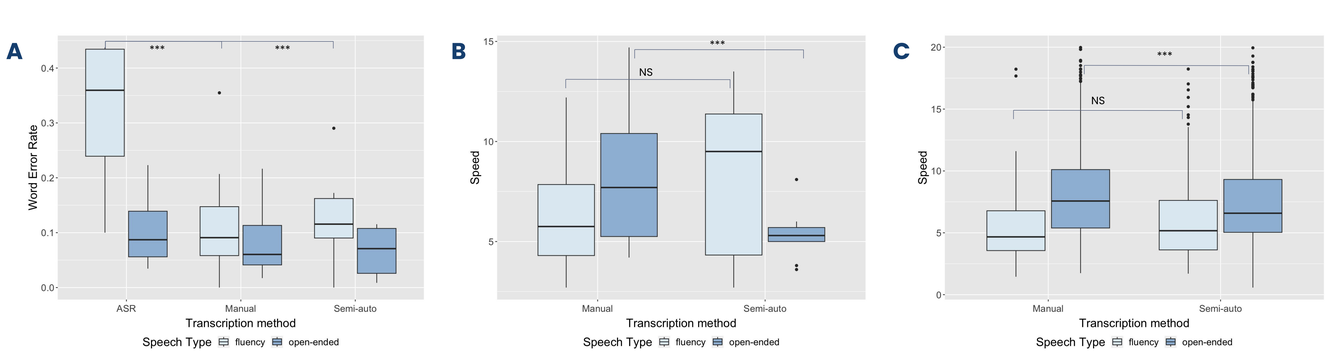


Figure 2A: Boxplot of word error rate (y-axis) by transcription method (x-axis). No significant differences were found except between ASR and both manual and semi-auto on fluency tasks. Figure 2B: Boxplot of transcription speed (the ratio of work duration to audio duration - lower is faster) and transcription method on the x-axis (ASR is excluded because the throughput of ASR is essentially 1 and not of interest in this study). While no significant difference was found between transcription methods for fluency tasks, semi-auto was significantly faster than manual transcription for open-ended tasks. Figure 2C: Boxplot representing transcription speed of the validation data set. Similar to the pilot data findings, transcription method did not impact speed for fluency tasks but did significantly speed up transcription of open-ended tasks.

Conclusion

Scalability and accuracy of transcription are essential to the future use of speech-based digital biomarkers in clinical development. The results here suggest:

- ASR has advanced significantly to reach comparable accuracy to human transcriptionists in open-ended speech tasks. While it cannot replicate all disease-relevant annotations or automatically derive quality metrics, it can be used on its own or in tandem with human review (semi-automatic) to significantly improve processing time and thus scale.
- For specialized samples, such as fluency or other constrained speech tasks, ASR produces significantly more errors than human transcriptionists. This highlights the developmental need for custom trained ASR models specifically optimized for each speech task type before the scaling benefits of ASR only or semi-automatic approaches can be realized.