Adjusting for Sex Improves the Screening Performance of a Digital Cognitive Assessment in an AD Memory Clinic Cohort

Johannes Tröger[1], Elisa Mallick[1], Stefanie Köhler[2], Alexandra König[1], Stefan Teipel[2], Nicklas Linz[1]

[1] ki elements GmbH, Saarbrücken, Germany [2]German Center for Neurodegenerative Diseases (DZNE), Rostock, Germany; Department of Psychosomatic and Psychotherapeutic Medicine, University of Rostock, Rostock, Germany

Introduction

Sex differences in cognitive performance are common across psychometric tests, clinical screening tools, and neuropsychological assessments, and can introduce systematic bias—especially in one-time, cross-sectional instruments used for Alzheimer's disease (AD) trial screening and dementia detection. Although some tests offer sex-specific norms, most do not, increasing the risk of unequal inclusion of men and women and reducing sample representativeness. Automatic digital cognitive assessments, including speech-based tools such as the ki:elements Speech Biomarker for Cognition (SB-C), may likewise show sex-related performance variance depending on task design and test paradigms. Prior work suggests sex effects in verbal measures, particularly verbal memory, with women often outperforming men, which could shift screening thresholds and impair classification accuracy.

Here, we examine whether sex adjustment improves the screening performance and inclusion balance of a digital cognitive assessment in an AD cohort.

Methods

We analyzed data from a memory-clinic cohort in the German DESCRIBE-AD registry (N=132; 64F; mean age 72.8 ± 7.6 years). Participants completed the ki:elements Speech Biomarker for Cognition (SB-C) digital screener alongside a standard neuropsychological battery and were clinically classified as healthy controls (HC; N=86; 44F) or cognitively impaired (CI; N=46; 20F).

SB-C scores were derived from recorded speech during Semantic Verbal Fluency (SVF) and the Rey Auditory Verbal Learning Test (RAVLT). Recordings were processed using ki:elements' proprietary pipeline (automatic speech recognition and multi-level feature extraction). The primary outcome was the SB-C global cognition score used in AD trial (pre-)screening.

We additionally computed two standardized scaled scores: (1) adjusted for age and education, and (2) further adjusted for biological sex. All score variants were developed and normed on an independent cohort; no fitting or calibration was performed on the DESCRIBE-AD sample.

Each variant was evaluated for discriminating HC vs CI by selecting an optimal cut-off (max F1), computing AUC, and deriving confusion matrices, with emphasis on whether sex adjustment improved performance and balanced classification outcomes across sexes.

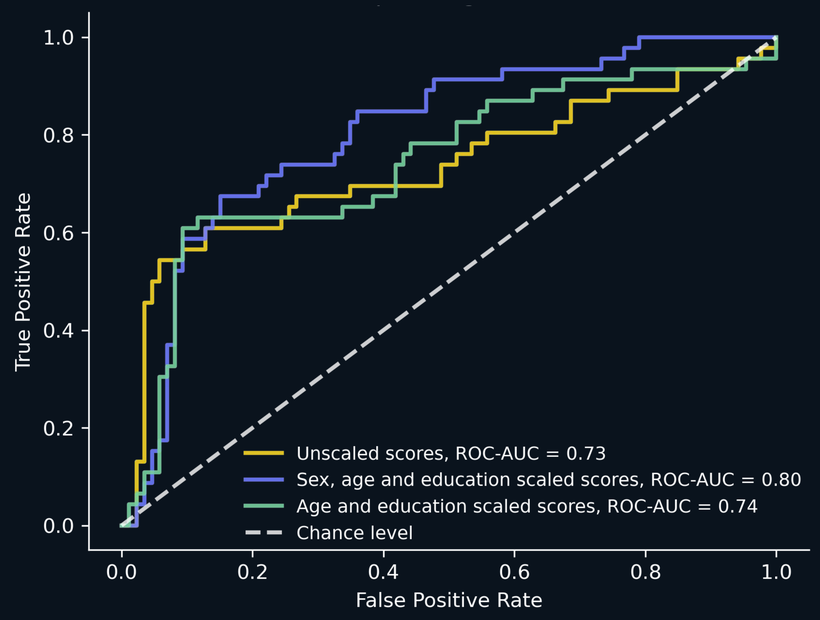

Figure: ROC-AUC depicting the screening performance of both sex-adjusted and non adjusted scores separating healthy controls from cognitively impaired individuals.

Table: Confusion matrix for screening performance of both scaled and the unscaled scores; HC=healthy controls & CI=cognitive impairment.

	Unscaled		Scaled for age + education		Scaled for age + education + sex	
	Correct: 49/64 F 52/68 M		Correct: 50/64 F 55/68 M		Correct: 48/64 F 56/68 M	
	0 HC	1 CI	0 HC	1 CI	0 HC	1 CI
0 HC	**33M / 40F**	9M / 4F	**34M / 42F**	8M / 2F	**35M / 38F**	7M / 6F
1 CI	7M / 11F	**19M / 9F**	5M / 12F	**21M / 8F**	5M / 10F	**21M / 10F**

Results

Screening performance improved when demographic adjustments were applied to the SB-C scores. The unscaled SB-C global score achieved an ROC–AUC of 0.73. Adjusting for age and education increased AUC to 0.74, and the inclusion of sex in the scaling model further improved performance to 0.80.

Examination of score distributions showed that, on average, women outperformed men on the SB-C, even after adjusting for age and education (SB-C_w > SB-C_m; $F = 17.0$, $p < 0.001$). Relative to their clinically established cognitive status, this pattern led to systematic misestimation: men's cognitive performance tended to be underestimated, while women's performance tended to be overestimated.

Confusion matrix analyses revealed that the sex-adjusted scaled score helps to account for these imbalances. Specifically, more men were correctly classified as HC and fewer were misclassified as CI following sex adjustment, which appears to drive the observed improvement in overall screening accuracy.

Conclusion

Overall, we present evidence showing that accounting for sex can improve the AUC screening performance and confusion matrix balance of a digital speech-based cognitive assessment in an AD memory-clinic cohort screening scenario. Failure to adjust for sex in a verbal neuropsychological test paradigm may lead to systematic misclassification while incorporating sex into the SB-C scaling procedure yielded more balanced classification outcomes and a marked increase in AUC. These findings highlight the importance of considering sex effects when screening with digital cognitive assessment that rely on a verbal test paradigm.

Disclosure

JT, EM, NL and AK are employed by the speech biomarker company ki:elements. JT and NL hold shares in the speech biomarker company ki:elements. The other authors have nothing to disclose.