Preprocessing variability as an analytical-validation dimension for speech-based Parkinson's disease endpoints

Submitter Lisa Chiang

Affiliation Cephalgo

SUBMISSION DETAILS

I agree to provide poster pdf for attendee download. Yes

What is the Methodological Issue Being Addressed? Speech-based digital endpoints are increasingly used in longitudinal Parkinson's disease (PD) trials, where the device is the patient's own smartphone and follow-up spans months to years. Devices receive routine operating-system (OS) updates that revise on-device noise cancellation (NC), unflagged and unlogged. For a patient whose clinical state is unchanged, the resulting shift in model output is indistinguishable from clinical change. Within the V3 framework this falls under analytical validation: whether output stays stable under realistic input variation. We ask whether preprocessing-induced drift can be quantified under controlled conditions, and whether feature-representation design changes how much drift an endpoint carries.

Introduction We describe a simulation-based strategy for quantifying preprocessing-induced drift and evaluate whether a multi-stream feature design reduces it. The design combines three streams (linguistic, acoustic, deep-learning) that differ in sensitivity to recording-channel change while each carrying disease signal, on the rationale that partially independent channel responses yield smaller net output shift when one channel property changes.

Methods Italian Parkinson's Voice and Speech Dataset: 28 PD, 22 age-matched controls. Subject-stratified 80/20 hold-out, no speaker overlap; the held-out test set comprises 9 unseen subjects (5 PD, 4 controls) contributing 136 recordings (72 PD, 64 control), all metrics on held-out subjects only. Worst-case specificity therefore rests on 4 control speakers. Validation against real OS-update pairs is methodologically confounded for isolating the preprocessing effect: because a past OS version cannot be re-run on a device, obtaining a real pair requires replaying each recording through a loudspeaker and re-capturing it on a handset, introducing a fresh, uncontrolled acoustic path per take. A preliminary two-version comparison showed exactly this confounding, with re-recording variance exceeding the preprocessing change. Simulation was therefore adopted to hold source audio fixed and vary only the denoising stage, using the most widely deployed algorithms, verified against reference iOS recordings by power spectral density. Each recording received a fixed level of additive Gaussian noise, held constant while denoised in turn by spectral subtraction (launch), Wiener filter (first update), and wavelet denoising (second update); only denoising differed. Models were trained on the launch configuration, or launch plus first update, and tested on a later unseen configuration across three out-of-domain conditions. Five architectures were compared under a common XGBoost classifier, differing only in feature representation: three self-supervised foundation models, a handcrafted baseline, and the multi-stream model. Primary outcome: per-recording reproducibility (ICC(A,1), bootstrap 95% CI); secondary: worst-case AUC,

sensitivity, specificity. Reporting follows TRIPOD+AI.

Results All five models discriminated well in-domain (raw-to-raw AUC >0.96), so the comparison isolates robustness with baseline ability matched. Under the denoiser change, worst-case discrimination separated sharply: the multi-stream model held AUC 0.927 with reproducibility ICC(A,1) 0.751 (NC shift 2%), whereas single-stream baselines fell to AUC 0.64-0.92 with far lower reproducibility (ICC 0.40-0.65; shift up to 22%). The handcrafted stream retained discrimination (AUC 0.92) but was least reproducible; self-supervised embeddings were stable but weak (AUC 0.64-0.72). For the multi-stream model, worst-case sensitivity was 97.2% and specificity 60.9%, the weakest dimension across models and resting on 4 controls.

Conclusion A change in on-device noise cancellation moved model output for recordings whose clinical state was unchanged, and the effect size depended on the feature representation. The multi-stream design preserved discrimination while reducing measurement-induced variance, through the combination of streams rather than a correction step. Discrimination shows the endpoint responds to disease; stability under preprocessing change shows insensitivity to the device; neither alone suffices. These findings derive from a single controlled cohort under simulated change and do not establish deployment robustness. Validation in larger multi-site cohorts spanning multiple devices, OS versions, and environments is required. We propose preprocessing robustness as an explicit dimension of analytical validation for speech-based digital endpoints.

Co-Authors

Adrián Ortiz[1], **Lisa Chiang[1]**

[1] Cephalgo

Keywords

Keywords
speech digital biomarker
Parkinson's disease
preprocessing robustness
analytical validation
DiME V3

Guidelines I have read and understand the Poster Guidelines

Disclosures Both authors are employees of CEPHALGO SAS, which develops speech-based digital biomarker technology. This work was funded by the European Innovation Council Accelerator (ADAPTE, Grant No. 190129251). One or more authors report potential conflicts which are described in the program.

Related Tables and Supporting Materials ISCTM_Supporting_Materials.pdf