DATA

The data used for our analysis came from STUDY. It consists of 22 columns. Two are identifiers to distinguish patient and specific trial. Two of the data columns are demographical data: age as an integer and binary representation of sex where 1 is female. Two columns are clinician measurements of symptoms and are considered source of truth. The remaining 16 columns encode jitter, shimmer, and 5 other numeric values which describe characteristics of voice thought to be related to Parkinson’s Disease symptom severity. These values are continuous positive real numbers. All values were normalized to [0..1]. The data contains approximately 200 trials per patient for 42 patients with no null values, 5875 total records.

The clinician measurements were gathered using the NIH endorsed process for measuring Parkinson’s symptoms, the Unified Parkinson's Disease Rating Scale (UPDRS). Trained and certified providers use its three general categories to produces continuous values conducive to normalization.

The special-built device used to analyze patient speech gathers 16 distinct features. Five describe Jitter and 6 describe Shimmer. These measure potentially interesting fundamental variations in speech patterns, frequency and amplitude, respectively. NHR and HNR are a ratio if noise to various tonal components. RPDE, DFA, and PPE are additional measurements which are also potentially interesting.

The data set was downloaded from <https://archive.ics.uci.edu/dataset/189/parkinsons+telemonitoring> in November of 2023. They were originally collected by Athanasios Tsanas, Max A. Little, Patrick E. McSharry, Lorraine O. Ramig for *Accurate telemonitoring of Parkinson's disease progression by non-invasive speech tests* published in 2009 by *IEEE Transactions on Biomedical Engineering*. Any use or transferal of the data should include the above citation and reference.