**PROBAST**

Study:

Comparing laboratory and in-the-wild data for continuous Parkinson's Disease tremor detection.

Step 2: Type of prediction study

**Is the study a diagnostic or a prognostic study?**

Diagnostic

**Is the study a development only, development and validation or validation only study?**

Development only

**What is the model of interest?**

Support vector machine

**What is the outcome of interest?**

PD tremor detection

Step 3: Assess risk of bias

**Domain 1: Participants**

**Describe the sources of data and criteria for participant selection**

Data were collected from 12 subjects (eight male, four female, ages 66 to 85) who had been diagnosed with PD two to five years prior. Each subject self-reported tremor in one or both hands.

**1.1 Were appropriate data sources used, e.g. cohort, RCT or nested case-control study data?**

Y

**1.2 Were all inclusions and exclusions of participants appropriate?**

Y

**Risk of bias introduced by selection of participants:**

Low

**Rationale of bias rating**

No further eligibility criteria.

**Domain 2: Predictors**

**List and describe predictors included in the final model, e.g. definition and timing of assessment**

Accelerometer data from wearable

**2.1 Were predictors defined and assessed in a similar way for all participants?**

Y

**2.2 Were predictor assessments made without knowledge of outcome data?**

Y

**2.3 Are all predictors available at the time the model intended to be used?**

Y

**Risk of bias introduced by predictors or their assessment**

Low

**Rationale of bias rating**

Accelerometer data are independent and applicable.

**Domain 3: Outcome**

**Describe the outcome, how it was defined and determined, and the time interval between predictor assessment and outcome determination:**

To improve label accuracy, subjects were only asked to record the amount of tremor they experienced within the 5 minutes prior to submitting the entry. Following the recommendation given in previous work [22], we used stratified rather than binary weak labels. That is, rather than asking subjects whether they experienced or did not experience tremor within the previous five minutes, we instead provided three label options (Almost none, Half the time, and Almost always). We chose to use three options, a slight deviation from the four strata used by Zhang et al. [22], because we felt that subjects would be able to more accurately select from a smaller set of options

**3.1 Was the outcome determined appropriately?**

Y

**3.2 Was a pre-specified or standard outcome definition used?**

Y

**3.3 Were predictors excluded from the outcome definition?**

Y

**3.4 Was the outcome defined and determined in a similar way for all participants?**

Y

**3.5 Was the outcome determined without knowledge of predictor information?**

Y

**3.6 Was the time interval between predictor assessment and outcome determination appropriate?**

Y

**Risk of bias introduced by the outcome or its determination**

Low

**Rationale of bias rating**

Patient reported outcomes are asked every hour and have to be filled in at the same time. Good standard for reporting of tremors for PD.

**Domain 4: Analysis**

**Describe number of participants, number of candidate predictors, outcome events and events per candidate predictor**

Six patients monitored for 2-4 weeks.

**Describe how the model was developed, predictor selection and risk group definition**

Using the WILD data from the test subject, we trained a stratified, Multiple Instance SVM (MI-SVM), as was used by Zhang et al. in [22]. We assigned approximate tremor percentages of [0-33%], [33-66%], and [66-100%] to the labels Almost none, Half the time, and Almost always, respectively.

**Describe whether and how the model was validated, either internally (cross validation, random split sample) or externally (e.g. temporal validation, geographical validation, different setting, different type of participants)**

LOOCV

**Describe the performance measures of the model, e.g. calibration, discrimination, classification, net benefit, and whether they were adjusted for optimism**

Mean absolute error for WILD

**Describe any participants who were excluded from the analysis**

None

**Describe missing data on predictors and outcomes as well as methods used for missing data**

Note that, for some partitions, training data would lack segments with“Almost none” or “Almost always” labels, making it not possible to initialize the stratified MI-SVM algorithm. Such partitions were ignored during model selection.

**4.1 Were there a reasonable number of participants with the outcome?**

N

**4.2 Were continuous and categorical predictors handled appropriately?**

PN

**4.3 Were all enrolled participants included in the analysis?**

Y

**4.4 Were participants with missing data handled appropriately?**

PN

**4.5 Was selection of predictors based on univariable analysis avoided?**

Y

**4.6 Were complexities in the data (e.g. censoring, competing risks, sampling of controls)**

**accounted for appropriately?**

Y

**4.7 Were relevant model performance measures evaluated appropriately?**

N

**4.8 Were model overfitting and optimism in model performance accounted for?**

Y

**4.9 Do predictors and their assigned weights in the final model correspond to the results**

**from multivariable analysis?**

U

**Risk of bias introduced by the analysis**

High

**Rationale of bias rating**

Low amount of patients and outcomes. Labels were first categorized (ordinal) but then MAE was taken so label was treated as continuous. Other metrics should have been reported e.g. accuracy, rmse etc. Missing data probably not handled correctly as some training data were simply removed.

**Overall Risk of bias**

High