**TADPOLE Submission - Methods description**

**Administration details:**

Team name: *BORREGOSTECMTY*

Type of team: *Regular*

Contact person: *José Gerardo Tamez-Peña*

Email address: *jose.tamezpena@itesm.mx*

Submission type: *full*

Name of the submission file (.xlsx): TADPOLE\_Submission\_BORREGOSTECMTY.xlsx

**Method category:**  
Ensemble of statistical predictions using regression

**Features:**  
1) Main cognitive tests (excluding subtypes), 2) MRI ROIs (Freesurfer), 3) APOE status, 4) Demographic information: age, gender, education and 5) Diagnosis status were used to build logistic statistical models for diagnosis forecasting and statistical regression models for ADAS13 and Ventricle Volume forecasting.

*Feature post-processing:*

The forecasting strategy derived a new set of features from the MRI set. It computed: the cubic root of all volumes, the square root of all surface areas, the compactness (), as well as the coefficient of variation () for all thickness measurements. Finally, the feature post-processing included the computation of the mean value and the absolute difference between left and right measurements.

**Feature selection:**

The bootstrapped stage-wise selection (B:SWiMS) procedure implemented in the FRESA.CAD R package automatically selected the features for the regression models. The “Prediction Method Details” section describes the procedure.

**Training data set:**

The D1 set was used to train the models for D2 prediction. D1 minus D3 subjects were used to train D3 prediction using a cross-validation strategy.

**Prediction data set:**

*Only D2 and D3 were predicted.*

**Confounder correction:**

The forecasting strategy adjusted for gender and intra-cranial-volume (ICV) differences between subjects. The approach selected the cognitively normal subjects from the D1-D3 set. Then a simple regression model stratified by gender tested if each feature was associated to the ICV. If the test resulted positive, then the predicted residual was used. Finally, the inverse-normal-transformation computed the z-value for all features using the gender-stratified distribution of the control population.

**Missing data:**

The training procedure used subjects with complete observations. The missing data of the test subjects were imputed using a nearest-neighborhood strategy. D2 imputation included all the D1 and D2 time-points. D3 imputation used D3 and D1-D3 subjects. In other words, imputation used historical and D3 concurrent data as prior information. The normalized L1 distance (*d*i) between observed features determined the neighborhood of each subject (Ni=*d*i<*T*), where *T* is a user-defined threshold. The median feature-value of all subjects in Ni estimated the missing value.

**Automatic: Yes/No**  
*Yes*

**Processing time:**  
Prediction is quite efficient. Subject adjustment and z-standardization require less than 1ms per subject. The LM and Logistic procedures are fast. It takes less than 1ms of user time to estimate each patient outcome

Training time, on the other hand, is very variable. A PC with an i7-2600 @3.4Mhz chip took 18 hours to train 50 models for ADAS13 prediction on MCI subjects. While training 25 models for the Time-to-MCI conversion took 15 min.

**Prediction method details:**

All predictions were based on the ensemble prediction of statistical regression models. The diagnosis forecast used logistic models for the long-term status and linear regression models for the time-to-event estimation. The Ventricle Volumes and ADAS13 forecasting used linear regression models to estimate the change from the last observed value. The best regression models were automatically found using the Bootstrap Stage-Wise Model Selection (B:SWiMS) strategy implemented in the FRESA.CAD R package[1]. In summary B:SWiMS selects a set of bootstrap samples (samples with replacement). Then a forward-selection (FS) procedure step-wise selects the top filtered feature that best describes the outcome. Once the FS procedure is completed, an update procedure uses the selection frequency to insert all the features that significantly improve the classification or regression accuracy. After that, a backward-elimination (BE) procedure explores the statistical significance of each feature. The statistical significance is adjusted for false discovery using the Benjamini-Hochberg approach, non-significant features are step-wise removed. The selected features are removed and the FS-update-BE stages are repeated until no more models can be found or the maximum numbers of models are reached (Ten for this experiment). Finally, a bagging procedure ensemble all the models into a final prediction model. The next paragraphs describe the details of the diagnosis and regression methods. The B:SWIMS procedure were run using the default FRESA.CAD parameters.

1. Subject classification method for forecasting cognitively normal (CN), mild cognitive impairment (MCI) or Alzheimer's disease (AD).

The diagnosis forecasting used a two-stage approach: Prognosis and Time-to-Event.

* 1. **Prognosis**: D1 subjects that undergo a diagnosis change were cases. D1 Subjects that remained stable for more than 5 years were controls for CN and 4 years for MCI. B:SWiMS was used inside a tenfold cross-validation (CV) for CN conversion or a fivefold CV for MCI conversion. At each fold, only one random visit per subject was used to build the logistic models (The random selection was done 5 times). D3 subjects that did not enter the training set were predicted; hence at the end of CV cycle, we will have a blind prediction for each D3 subject. The training approach repeated the procedure 5 times for a total of 50/25 predicting models. The median value of the 50/25 predictions per D2 subject was used as the final prediction. The median of either the 10/5 predictions (Subjects that were used in training) or the 50/25 predictions for subjects that never were used for training was used for the final D3 prediction.
  2. **The time-to-event:** The procedure used a 5 fold CV procedure, but instead of estimating the future outcome, it estimated the square root of the time it took a case subject to convert from its current status to AD or MCI. The continuous distribution function (CDF) of the 25 time-to-event estimations was evenly sampled from January of 2018 to Dec 2022. The forecast consisted in multiplying the probability of diagnosis conversion by the value of time-to-event CDF.
  3. Reported p-values. The final prediction took into consideration the models for CN-to-MCI, CN-to-AD, and MCI-to-AD. A subject was not allowed to change from MCI-to-CN nor AD-to-MCI. The final know diagnosis determined which model to be used for the final prediction. The CN probability was the average p-value of the CN-to-MCI and CN-to-AD on subjects that started from a CN status.

1. Forecasting ventricle volume and ADAS13 score
   1. In a similar fashion to the construction of the logistic models, the forecasting of the continuous variable used linear models. A five-fold CV procedure, repeated 10 times, of the B:SWiMS procedure estimated 50 linear forecasting models. At each interaction, the differences in the continuous feature to the last observed value of the feature were estimated using the time-interval (t), the t2, log (t) and second order interaction of log (t) to the top discriminant feature.
   2. The forecasting approach computed change models for CN subjects, MCI subjects, and AD subjects. The final know diagnosis of the subject determined which model to be used for the final forecast. The time from this last observation to future prognosis date yield the t. After forecasting the change, the procedure added the last know value of the feature. The mean value of the 50 estimations for each D2 subject and their corresponding empirical 50% confidence intervals were reported. For D3 subjects either the 10 blind estimations or the 50 estimations determined the final reported value.

1. Tamez-Pena JG, Martinez-Torteya A, and Alanis I. Package ‘FRESA. CAD’. 2014.