1. Background

The relapse remitting multiple sclerosis (RRMS) treatment landscape has undergone considerable changes since the emergence of interferon-β (Betaseron, Rebif, Avonex, Extavia) and glatiramer acetate (Copaxone) in the mid-1990s (these two product classes will herein be referred to as BRACE therapies). The increased availability of disease modifying therapies (DMTs) for the treatment of RRMS has increased the complexity of optimizing treatment choices for practicing neurologists. It is important to understand whether historical patient data can be used to predict future disease activity for RRMS patients and whether this can be used in clinical practice to improve patient outcomes.

1. Objectives
   1. The primary objective of this study is to estimate the overall probability of relapse over a twelve-month period
   2. The secondary objectives of this analysis are to produce predictions of disease activity and progression over a twelve-month period as defined by a number of outcome measures. These outcome measures are: probability of EDSS progression, probability of confirmed EDSS progression, probability of relapse or EDSS progression, probability of relapse or confirmed EDSS progression, and probability of relapse and EDSS progression. Additional secondary objectives will estimate all of the above outcome measures under a number of different treatment scenarios, e.g. predictions for outcomes will be made for cohorts switching between pre-defined treatments.
   3. Further, exploratory objectives produce predictions using advanced machine learning methods to compare these with baseline estimates using penalized multiple logistic regressions
2. Study design

To address the objectives of this study, a retrospective cohort analysis will be conducted using electronic medical record (EMR) data extracted from the NeuroTransData (NTD) network of neurology practices in Germany. Data will be extracted for RRMS patients receiving BRACE therapy and will be divided into **five cohorts**:

* 1. Patients continuing on BRACE therapy
  2. Patients switching to alternative BRACE therapy
  3. Patients switching to first-line oral therapy
  4. Patients switching/escalating to second-line therapy
  5. Composite cohort combining the above cohorts

The index date is defined as the date of

* + - * 1. Switch/escalation
        2. The most recent visit that will allow for 12 months of post-index data (in the BRACE continuation cohort)

The analysis will model six alternative, binary outcome measures:

1. Relapse
2. Expanded Disability Scale (EDSS) progression
3. Confirmed EDSS progression
4. Relapse or EDSS progression
5. Relapse or confirmed EDSS progression
6. Relapse and EDSS progression
7. Create cohort data
   1. Cut the continuous variables into buckets
   2. For categorical variables:
      1. For those variables whose levels have ordinal relationship, merge small levels in categorical variables into 1 level with order (e.g. edss\_score)
      2. For those variables whose levels have no ordinal relationship, merge small levels in categorical variables into 1 level without order (e.g. birth\_region)
   3. Convert NA to 999
   4. Convert categorical and numeric variables to dummy variables
   5. Do some adjustment according to client’s requirement for the variables included in modeling
8. Create descriptive stats table
9. Create cohort data for modelling
   1. Start from the data in section 4.e.
   2. Remove reference categories
10. Modelling
    1. Elastic-net with all the base variables
    2. Elastic-net with the top 10 important variables
    3. Standard logistic regression with the top 10 important variables
11. Information we focus on for the above three type models
    1. Model performance
       1. AUC
       2. Corresponding confidence interval
    2. Variable importance
       1. Number of Times Variable Retained
       2. Average Coefficient
       3. Average Odds Ratio
    3. Odds ratio for standard (unconstrained) logistic regression based on most important ten variables
       1. Odds ratio
       2. Corresponding confidence interval
       3. P-value
    4. Actual outcomes by quintile of predicted risk score based on logistic regression with elastic-net for most important ten variables
       1. 5 quintile Groups
12. Cross-validation resampling

At each step that splits patients into folds/samples, stratified random sampling will be applied to ensure that the same proportion of positive outcomes (i.e. actual positives for the outcome of disease activity) are in each fold.

For each of the five outcomes of disease activity, two nested loops will be applied within each of the five cohorts to compute the average AUC and average variable rank. The steps are described below

* 1. Randomly stratify each cohort into ten evaluation folds / samples of equal size.
  2. Assign one evaluation fold as the test set and the remaining nine evaluation folds as the training + validation set.
  3. Further randomly stratify the training + validation set into ten validation folds of equal size.
  4. For every combination of possible values for hyper-parameters α and λ:
     1. Assign one validation fold as the hyper-parameter validation set and the remaining nine validation folds as the hyper-parameter training set;
     2. Estimate a logistic regression with an elastic-net penalty using the hyper-parameter training set (i.e., the nine out of ten validation folds);
     3. Compute the AUC for the hyper-parameter validation set (i.e., the tenth left-out validation fold);
     4. Repeat steps 4 a-c for all ten validation folds;
     5. Compute the average AUC for the ten validation folds (i.e., across all ten validation folds);
     6. Repeat steps 4 a–e for all possible hyper-parameter combinations.
  5. Select the hyper-parameter combination that produces the highest average AUC.
  6. Use the selected hyper-parameter combination to estimate a logistic regression with an elastic-net penalty on the training + validation set (i.e., the nine evaluation folds).
  7. Compute the AUC for the tenth evaluation fold using the fitted model in step f,
  8. Compute the importance ranking of variables for the tenth evaluation fold using the fitted model in step f,
  9. Repeat steps b-h for all ten evaluation folds.
  10. Compute the average AUC and variable importance (i.e. across all ten evaluation folds).
  11. Select a pre-defined number of most important variables based on the results obtained in step j.
  12. Compute a logistic regression without regularisation using the selected most important variables.

1. Variable exploration
   * 1. Use Random Forest instead of elastic-net
     2. Divide all other predictors in the raw data into groups. For how to define the groups (e.g., visit reasons is one group, Cerebrospinal fluid is another, etc).
     3. Cohort + outcome: Focus on the composite cohort + 3 types of outcomes, namely relapse, edss progression and confirmed edss progression.
     4. Data size: stratify the data into 2 halves (different for every outcome). Hide one half permanently (we’ll probably never touch it). Then we use the other half to do the following experiments.
     5. Say there are K groups obtained from Step 10.ii.. Then we do 2 \* K experiments for each of the cohort + outcome combinations.
   1. For each experiment, use all the base predictors + ONE group of new variables, produce the test and training AUCs (Remember that everything in this experiment only uses one half data produced in Step 10.iv. above!!)
   2. For each specific new group, do two experiments: one using our existing elastic net and the other using the newly implemented random forest. This will give us two test AUCs for the same group, which tells us whether the new group is helping the prediction.
   3. We do this for all K groups. This tells us which groups are helping.