

# Modelling the MESS Data

*Suzie Brown and Marco Palma*

*1 November 2017*

## Introduction

The MESS data comes from a randomised controlled trial about treatments for epilepsy. The subjects in the study are patients who have had only one seizure or have not had epilepsy for long. There are two treatments; either immediate or deferred. Patients assigned the deferred treatment are not prescribed anti-epileptic drugs after their first seizure, but the decision is revisited if they have further seizures.

Because many people have only one seizure in their life, we don't necessarily want to put patients on anti-epileptic drugs after just one seizure. However, we also don't want patients going on to have further seizures that could have been prevented or reduced by starting treatment immediately.

We are therefore interested in the level of benefit associated with immediate treatment relative to deferred treatment. Ultimately we could weigh this up against the costs of ongoing medication in order to decide whether a new patient should be given immediate treatment or not. This decision could depend on other data about the patient; for instance the trial recorded demographic details, results of some medical tests, and information about previous seizures.

Our aim is to construct a model considering some of these factors, which can be used to predict the outcome for a specific type of patient under immediate and deferred treatment. The model could then be used with some loss function to make optimal decisions about how to treat new patients.

## Exploratory analysis

### Change data types

Treatment 0/1 and sex 0/1 have been stored as numeric; we change the data type to factor.

We see that there are many NAs in the columns relating to dates, but this is not shown on the summary because they are formatted as character. We format the dates as dates.

We assume the `centre` variable must be an identifying number for the clinic where the subject entered the trial. As such, we format it as a factor.

We find that the binary results for the various medical tests have been stored as 1/2 = yes/no. We convert these to 0/1 = no/yes.

### Check censoring indicator

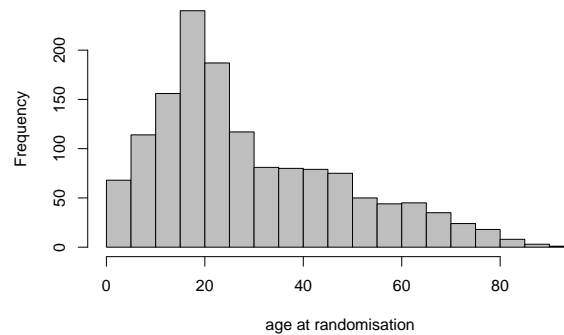
Next we want to see which way round the censoring indicator is used. We see that where the censoring indicator `ind1yr` is 1, the time to one year remission `int1yr` has minimum 365, whereas where `ind1yr` is 0, the minimum is 1. From this we deduce that the indicator is 1 when the variable is observed and 0 if the variable is censored.

## Missing data

We verify that the same data are missing from `d1seiz` and `period`. These were probably subjects who couldn't remember the date of their first seizure, and/or whose medical records were missing. This suggests they are missing not at random - for instance if the subject can't remember the date it is likely to be a long time ago, i.e. higher values of `period`. However, since they are only 5 out of 1425 observations, we suspect it will not make much difference to treat them as if missing at random.

## Investigate some variables

Plotting the age at randomisation `ager` we see that it is positively skewed. This is plausible since it is a study of people with single seizure or recent diagnosis of epilepsy, and it is more likely that an individual has their first seizure at a younger age.



We see that the number of subjects having each treatment is roughly equal, in accordance with the design of the study.

```
## treatment
##    0    1
## 712 713
```

There is a mystery as to how some patients who have not had an EEG have recorded an abnormal EEG result. There are several possible explanations (different types of recording error), but we will assume the patients with an abnormal EEG have had an EEG. We adjust the `eeg` variable accordingly. Since there are only five subjects in this category it shouldn't substantially affect any results.

```
##    abEEG
## EEG    0    1
##    0 80    5
##    1 554 786
```

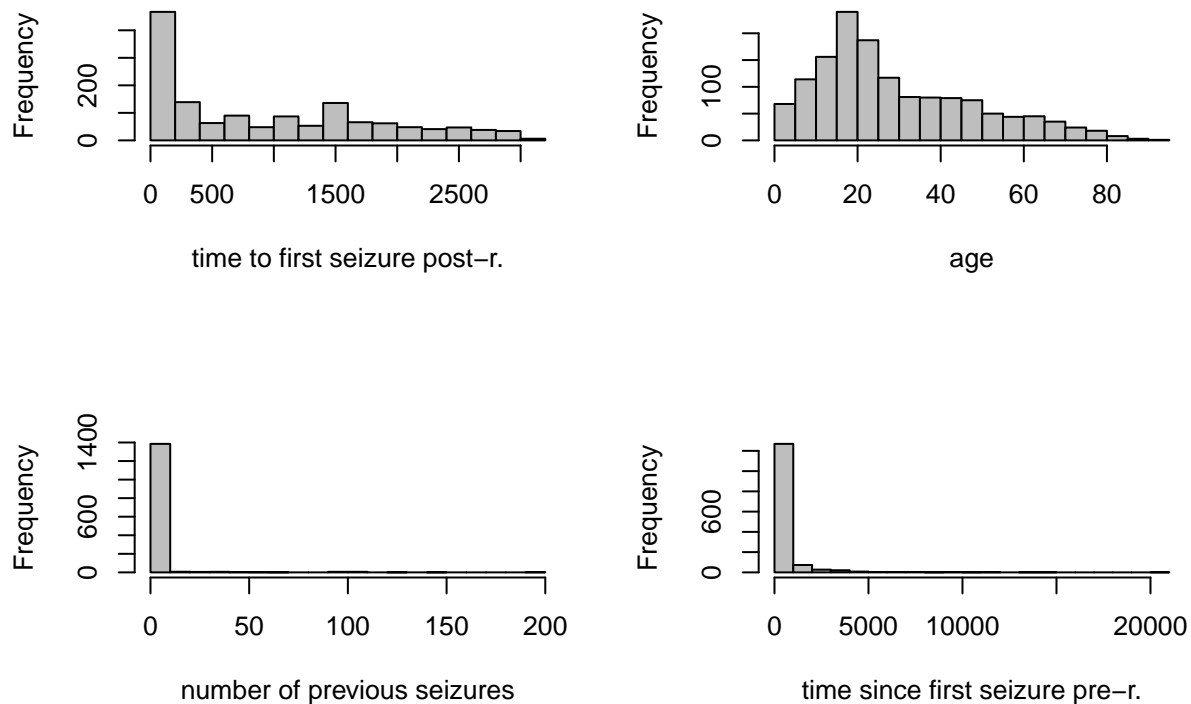
## Split age into bins

We use the age categories referred to in Marson et al. (2005).

```
## agecategory
## [0,5] (5,9] (9,19] (19,29] (29,39] (39,49] (49,59] (59,69] (69,92]
##    68    90   379   327   166   150   105    72    68
```

## Transform some variables

We plotted histograms of each of the key (non-binary) variables. The variables `nseiz` and `period` are very positively-skewed. This is not surprising since this is a study on early epilepsy, so most subjects have only had one seizure and the first seizure was not long before entering the study.



We will try transforming `period` and `nseiz`. For `nseiz`, we simply use the indicator of whether it is 1 or more than 1, since over half of the patients have `nseiz==1`. We take the logarithm of `period` to reduce the skew.

## Select initial set of covariates

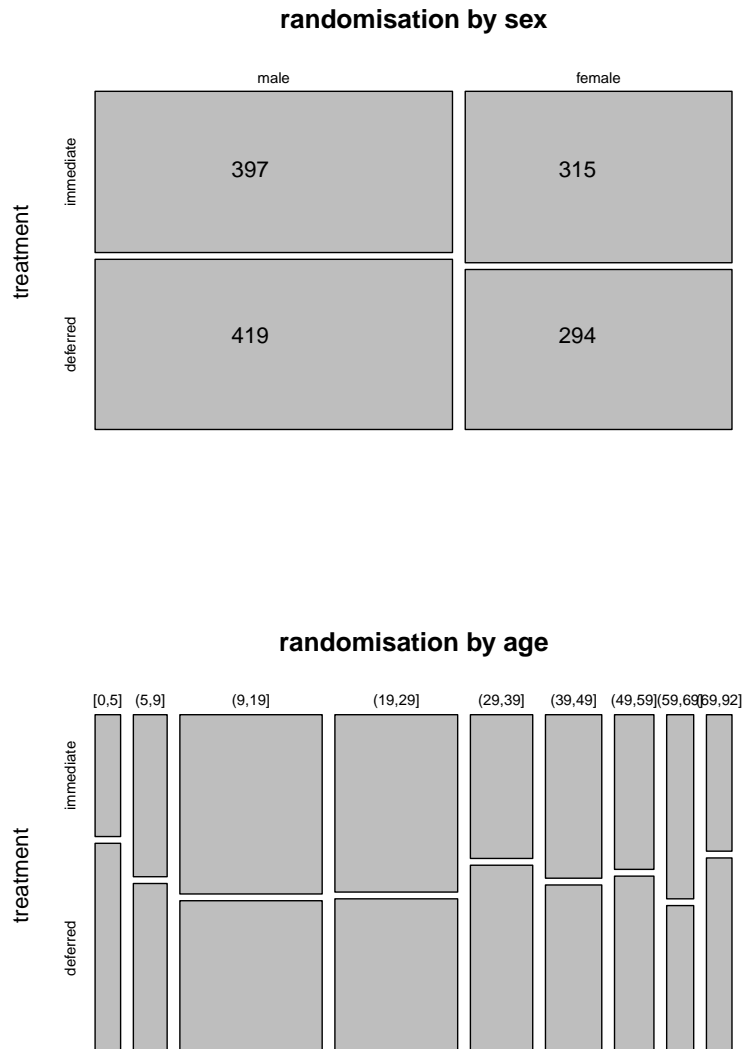
For the purposes of the initial exploratory model fitting, we use a small subset of the covariates. We address the issue of variable selection later on.

We drop all of the date variables because we suspect the absolute time does not matter; relative times are available for many of the events. We also drop the sub-categories for different types of seizures, retaining just the number of tonic-clonic seizures (the most severe type) and the overall number of seizures. We also drop the sub-categories of EEG abnormality, retaining just the overall indicator. We do not include the centre, since there are so many levels it adds too many degrees of freedom, and we have no information about how they might be grouped e.g. geographically.

```
covars1 <- transmute(mess, ager, sex, trt, ntc, nseiz.tr, period.tr, eeg, abeeg)
```

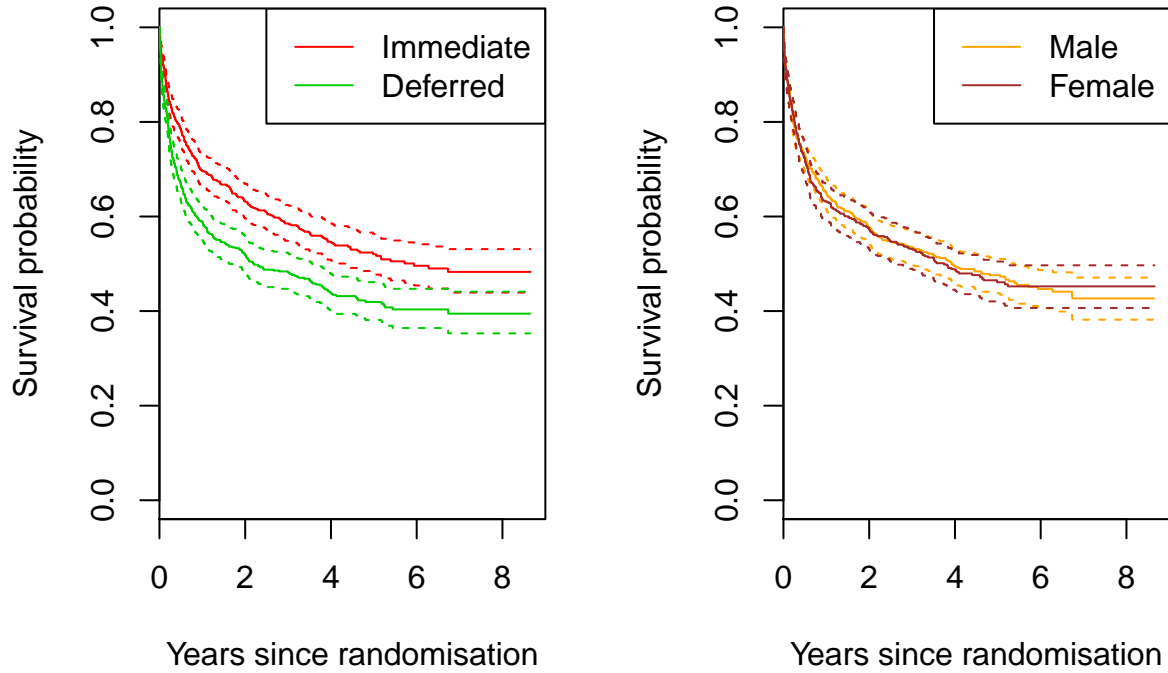
## Check balance of randomisation

We check that the randomisation is balanced over sex and age. There is a noticeable deviation from an even split in some categories, but the deviation is not systematic and the overall sample size is large so it should not make a big difference to our results.



## Exploratory Kaplan-Meier curves

We plot Kaplan-Meier curves of the first seizure after randomisation, grouped by treatment and sex. The first plot shows that the survival probability for those who received a deferred treatment is lower than for those immediately treated, at least for the first 5 years after the randomisation. From the second plot it seems that sex does not have a significant effect on the survival probability.



## Cox proportional hazards model

The Cox model relies on the strong assumption of *proportional hazards*, that is that a unit increase in a certain covariate has a multiplicative effect on the hazard rate. In particular, the hazard function  $\lambda$  has the form

$$\lambda(t|X) = \lambda_0(t) \exp(X^T \beta)$$

where  $X$  is the matrix of covariates and  $\beta$  is the vector of coefficients to be estimated, and  $\lambda_0(t)$  is the baseline hazard (i.e. when all covariates are zero). The exponential of a coefficient  $\exp(\beta_i)$  represents the multiplicative effect on the hazard resulting from a unit increase in the corresponding covariate.

First we fit a naive proportional hazards model with the covariates of interest. We see that the treatment `trt` is significant, as is the number `nseiz` of seizures prior to entering the study. This is to be expected since subjects with a tendency for more frequent seizures are likely to go less time before the first post-randomisation seizure regardless of treatment. The elapsed time `period` between the subject's first seizure and their entry into the trial is slightly significant, which is also unsurprising.

```
##           coef exp(coef)    Pr(>|z|)
## trt1      0.3262837253 1.3858085 2.141267e-05
## ager     -0.0032234409 0.9967817 1.285603e-01
## sex1      0.0129018821 1.0129855 8.686761e-01
## nseiz     0.0153455343 1.0154639 0.000000e+00
## period   -0.0000646568 0.9999353 3.965425e-02
```

Next we assess whether proportional hazards is a suitable assumption for these data. We see from the residual plots that the treatment seems to have a decreasing effect over time, but the residuals of the other covariates

don't seem to vary over time. This is confirmed by the small p-value for `trt` in the test of the proportional hazards assumption. The treatment effect decreasing over time is indicative that a proportional hazards model is not suitable, but we will try a few fixes before abandoning the Cox model completely.

Next we try adding into the model an interaction term of treatment with time, to correct for the problem with the first model. We see, as expected, that the interaction term is very significant. However the proportional hazards model is still rejected.

```
##               coef exp(coef)    Pr(>|z|)
## trt1          3.344520e-02 1.0340108 0.74642186
## age           -2.549841e-03 0.9974534 0.24916927
## sex1          4.138589e-02 1.0422542 0.63117043
## nseiz         4.084081e-03 1.0040924 0.05821305
## period       -1.902268e-05 0.9999810 0.62300420
## trt0:intseiz1 -3.120390e-01 0.7319530 0.00000000
## trt1:intseiz1 -3.122835e-01 0.7317741 0.00000000
```

Next we fit the model using the transformed versions of `period` and `nseiz`. The significant covariates are largely the same as in the first model, however the assumption of proportional hazards is not supported in the `trt` and `nseiz` terms.

```
##               coef exp(coef)    Pr(>|z|)
## trt1          0.350541667 1.4198364 5.113729e-06
## age          -0.002726502 0.9972772 1.928886e-01
## sex1         -0.040488044 0.9603206 6.044246e-01
## nseiz.trTRUE  0.632578980 1.8824591 1.811173e-11
## period.tr    -0.044293505 0.9566731 4.568768e-02
```

Finally, in the next model we include an interaction term between `trt` and `nseiz`, since it is plausible that the treatment may have a different effect on people who suffer more severely (indicated by having had more seizures previously). With this model the proportional hazards assumption appears to be satisfied. However we find that the interaction term is only slightly significant, and including the interaction drastically reduces the estimated treatment effect.

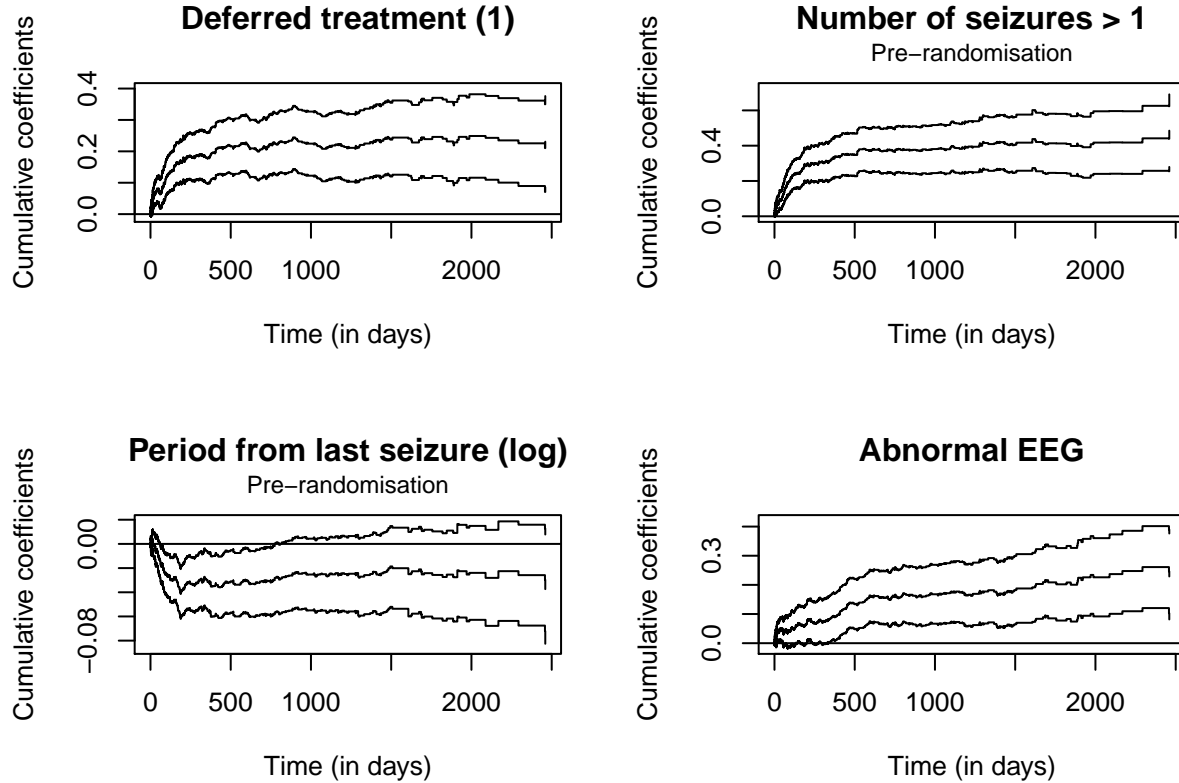
```
##               coef exp(coef)    Pr(>|z|)
## trt1          0.218269768 1.2439226 4.712228e-02
## age          -0.002818326 0.9971856 1.779277e-01
## sex1         -0.045140985 0.9558627 5.637268e-01
## nseiz.trTRUE  0.493678735 1.6383321 8.247602e-05
## period.tr    -0.044871490 0.9561203 4.269108e-02
## trt1:nseiz.trTRUE 0.256384008 1.2922489 9.468218e-02
```

## Aalen's additive model

The proportional hazards assumption underlying Cox model might be in many cases too strict and lead to unrealistic conclusions. Moreover, it is not uncommon to consider that the effect of one or more covariates might change over time. In these cases the model proposed by Aalen (1989) provides a flexible way to tackle these issues. Unlike Cox model, in Aalen's model the effect of the covariate on the hazard function at time  $t$  is additive. In other words, the hazard function at time  $t$  is a linear function of the covariates. Each coefficient is a function allowed to vary over time; in this sense, the model is fully nonparametric. At time  $t$ , the value of the coefficient function is to be read as the variation with respect to the baseline hazard function induced by a change of one unit (level in case of categorical variables) in the covariates.

Aalen's additive model is widely used because of the easiness in the interpretation of the graphical results, that report the cumulative regression coefficient over time with 95% confidence bands. The slope of the cumulative coefficient function gives indeed some information about the effect of a variable on the outcome of

interest. For what concerns the effect of deferring the treatment, the cumulative coefficient increases in the first two years approximately after randomisation, then remains flat: this suggests that the hazard ratio is fairly stable until that breakpoint and then it gets closer to 1. In other words, after two years having received the treatment immediately or not has no effect on the time to the next seizure. A similar conclusion can be drawn on the number of seizures pre-randomisation. In addition, the time from the last seizure . . . . . has not a significant effect after approximately 2 years, while the effect of an abnormal EEG is quite constant over time



## Parametric survival models

An alternative to the Cox model is to consider a parametric specification for the distribution of survival times. In particular, the class of accelerated failure time (AFT) models provides the natural extension in case the proportional hazard assumption does not work. Following the introduction provided by Hosmer, Lemeshow, and May (2008), it is possible to write the distribution of the time to event for the  $i$ -th subject as

$$\log(t_i) = \mu + \beta^T x_i + \sigma \epsilon_i$$

where  $\sigma$  is a scale parameter and  $\epsilon_i$  is an error term with a prespecified distribution (the choice for this distribution determines the type of model). Common choices for the regression models are exponential, Weibull, lognormal (when the distribution of  $\epsilon$  and therefore the distribution of  $\ln(t)$  is normal) and loglogistic: this choice will affect the shape of the hazard function, that may be monotonic or having a maximum dependent on the scale parameter. The effects in this model are multiplicative on the time scale: in other words, the time to the event depends on the exponential of the linear combination of the covariates weighted by the coefficients. The name “accelerated failure time” arises from the fact that given

$$t_i = \exp[\beta^T x_i] \exp[\sigma \epsilon_i]$$

the effect of the covariates is to “accelerate” or decelerate the time to the event of interest with respect to a subject with baseline characteristics (i.e., the one for which all the covariates assume value zero). This means that if for a given variable the coefficient is positive, the time to the event will increase, hence the effect of the covariate will be protective. Therefore, the sign of the coefficient has to be read in opposite way with respect to the proportional hazards models.

Among the accelerated failure time (AFT) models, lognormal model is preferred.

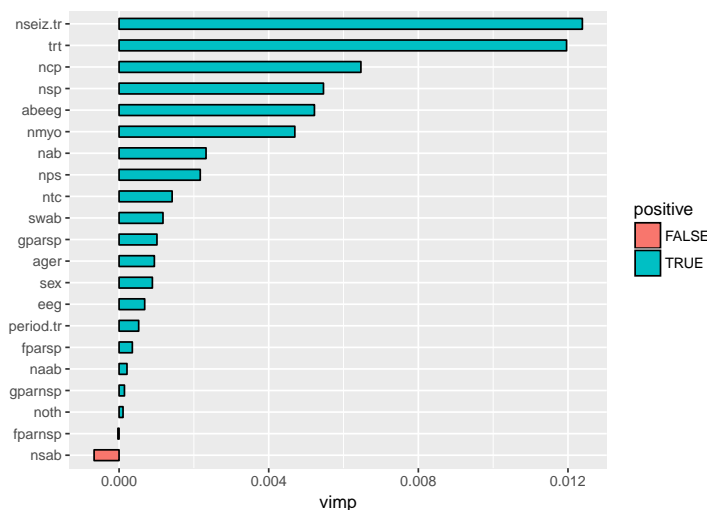
##		N.obs	Loglik	df	AIC
##	exponential	1420	-5900.726	7	11692.63
##	weibull	1420	-5548.895	8	11024.34
##	lognormal	1420	-5511.851	8	10950.74
##	loglogistic	1420	-5527.916	8	10979.20

## Random survival forests

We would like to ensure that the set of covariates we include in the model is relatively stable. Up to now we have only selected covariates insofar as the fitted log-normal model has some significant and some insignificant covariates. We now compare the set of significant covariates in the log-normal model to those obtained using a non-parametric variable selection technique, namely random survival forests.

Again we omit the **centre** variable since it destabilises the model, having a large number of levels many of which correspond to only one observation. If we had some additional information say about suitably groupings of centres it would be sensible to include it. For instance, in Marson et al. (2005) the authors distinguish between UK and non-UK centres. However we do not have any information linking the number of a centre to its location. However we now include the counts of every type of seizure, and the results of all tests, as possible covariates.

We still omit date variables. we found that the error rate didn’t really decrease by creating more than 100-200 trees, so we use **ntree** = 200. There is not a universal threshold on the **vimp** above which variables should be considered selected, but in the literature **vimp**=0.02 has been suggested as a rule-of-thumb.



As we found in our exploratory analysis, **nseiz** and **trt** are the most important covariates, with **abeeg** also significant. We also see the counts of several particular types of seizure are important, which weren’t included in our initial analysis.

The six most important variables in random forests were all significant in the log-normal model, which is encouraging. On the whole, random forests has selected roughly the same variables as the log-normal model,



with only a few discrepancies. For instance, **nps** (one of the seizure counts) was significant in the log-normal model ( $p=0.009$ ) but was not selected by random forests.

For the purposes of our model, we will include only the covariates which the parametric and non-parametric approaches have both selected, namely **nseiz**, **trt**, **nep**, **nsp**, **abeeg**, and **nmyo**. This captures the most important covariates of each approach.

## Further research

## References

- Aalen, Odd O. 1989. "A Linear Regression Model for the Analysis of Life Times." *Statistics in Medicine* 8 (8). Wiley Online Library: 907–25.
- Hosmer, David W., Stanley Lemeshow, and Susanne May. 2008. "Parametric Regression Models." In *Applied Survival Analysis*, 244–85. John Wiley & Sons, Inc. doi:10.1002/9780470258019.ch8.
- Marson, A, A Jacoby, A Johnson, L Kim, C Gamble, D Chadwick, Medical Research Council MESS Study Group, and others. 2005. "Immediate Versus Deferred Antiepileptic Drug Treatment for Early Epilepsy and Single Seizures: A Randomised Controlled Trial." *The Lancet* 365 (9476). Elsevier: 2007–13.