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 were recently diagnosed. There are two treatments, immediate and 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 information about the patient; for instance demographic details, results of 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

We first format as factors those variables that should be factors, format the dates as dates, and change the 1/2-coded binary variables to 0/1.

Check censoring indicator

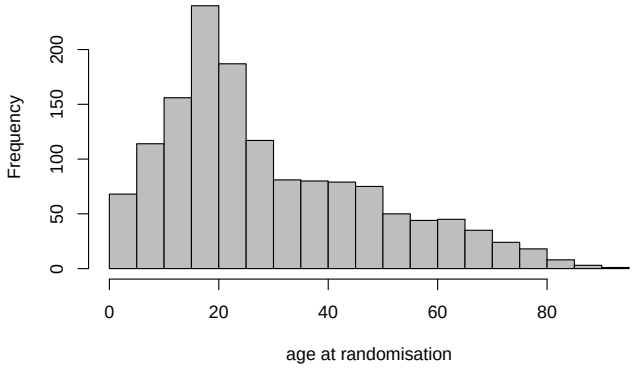
Next we want to see which way round the censoring indicator is used. We see that where the censoring indicator `ind1yr` is 1, the time to one year remission `int1yr` has minimum 365, whereas where `ind1yr` is 0, the minimum is 1. From this we deduce that the indicator is 1 when the variable is observed 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 assume 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 and early epilepsy, and it is 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
## Immediate  Deferred
##       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` varaible accordingly. Since there are only five subjects in this category it shoudln't substantially affect any results.

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

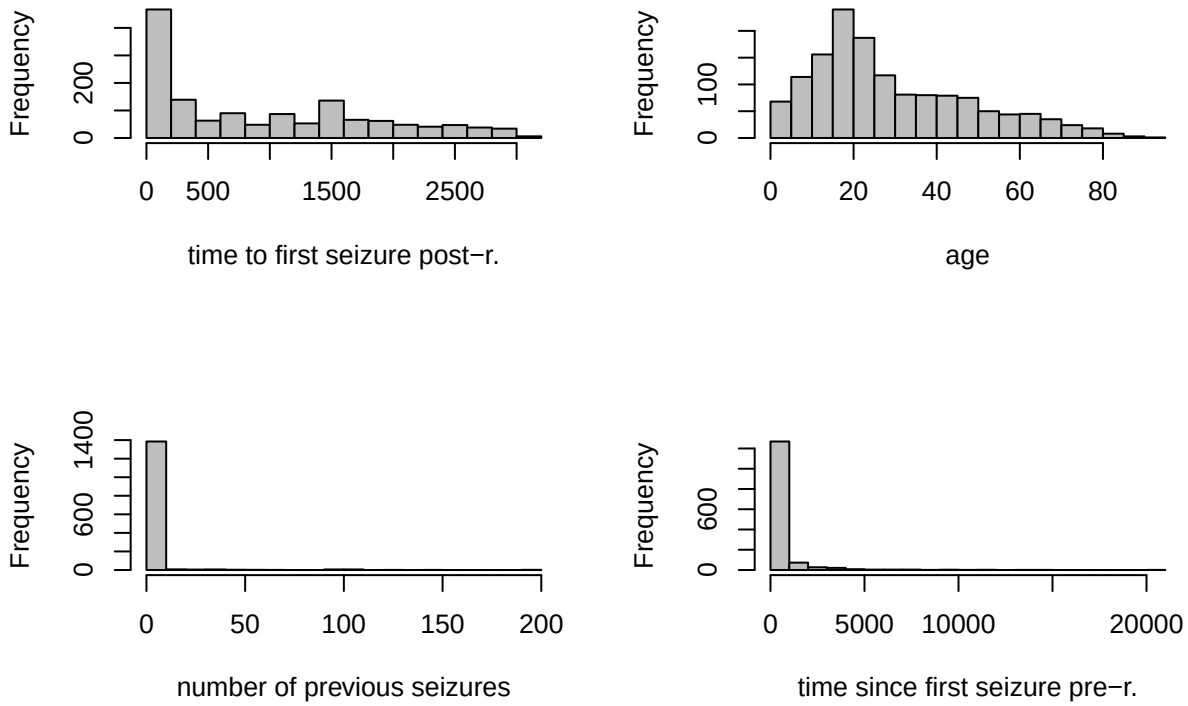
Split age into bins

We use the age categories referred to in Marson et al. (2005). The table shows the number of subjects in each category.

```
## 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 apply transformations to `period` and `nseiz`. For `nseiz`, we simply use the indicator of whether it is 1 or greater than 1, since over half of the patients have `nseiz=1`. For `period` we take the logarithm to reduce the skew.

Select initial set of covariates

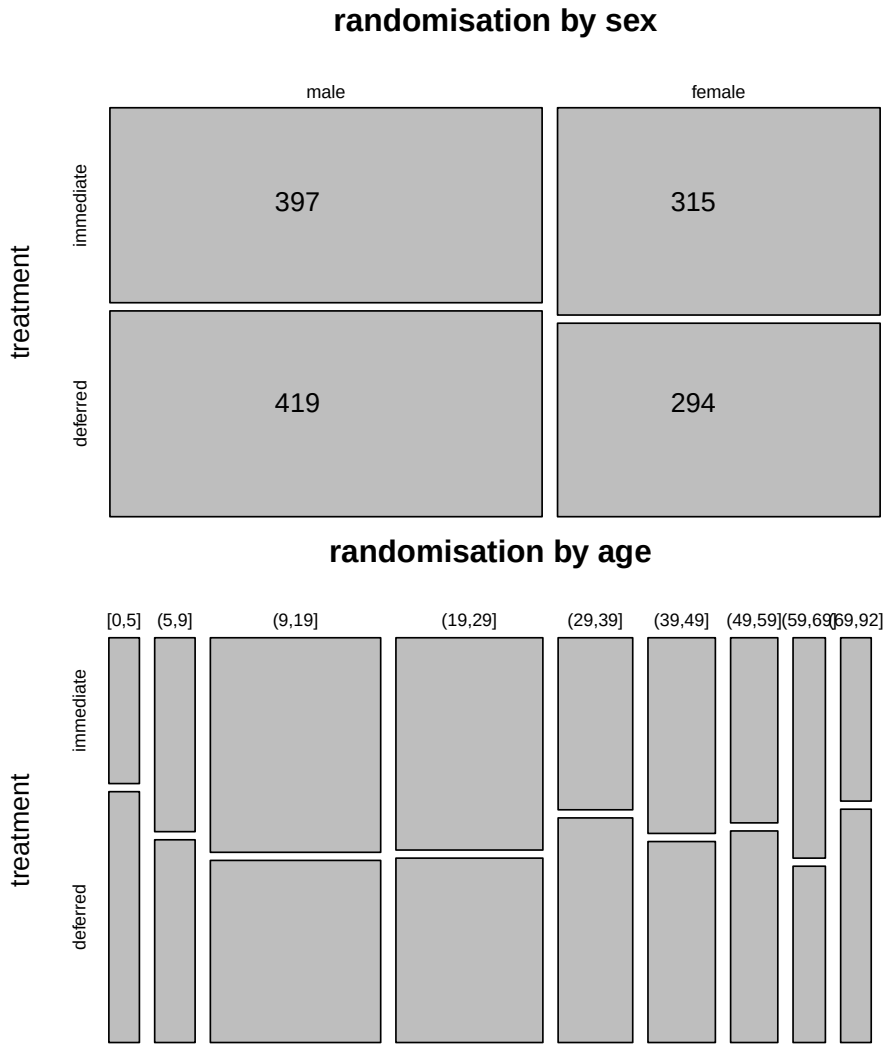
For the purposes of the initial exploratory model fitting, we use a small subset of the covariates. We will address the issue of variable selection later on.

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

We drop all of the date variables because we suspect that absolute times do not matter: relative times are available for the main 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.

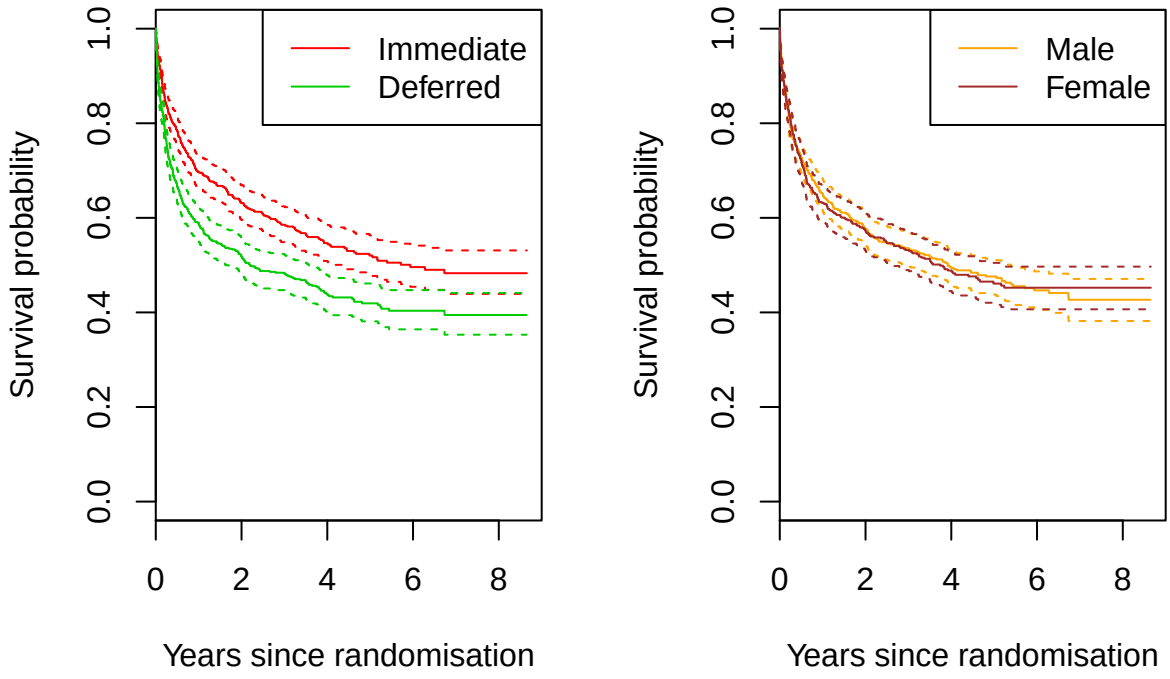
Check balance of randomisation

Next 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 significantly bias our results.



Exploratory Kaplan-Meier curves

We plot Kaplan-Meier curves of the time to 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 λ has the form

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

where X is the matrix of covariates, β 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.

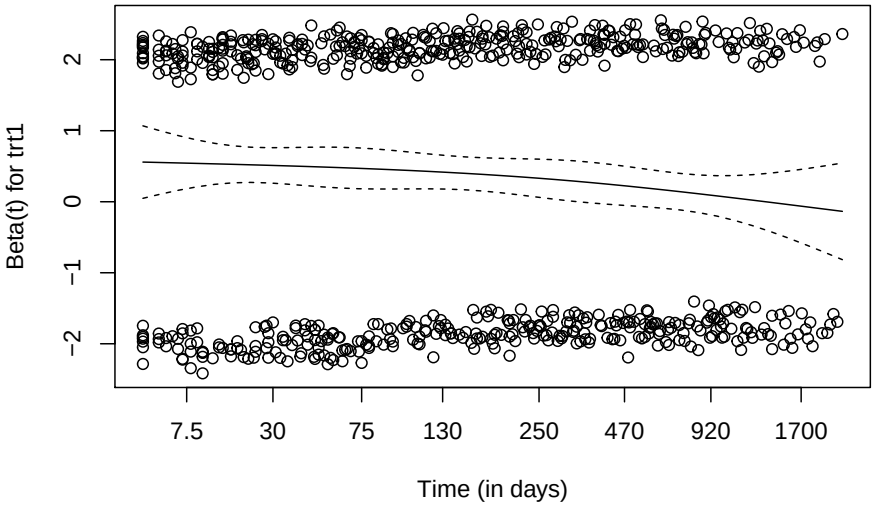
First we fit a naive proportional hazards model with the covariates of interest. We see that the treatment `trt` is significant, as is the number `nseiz` of seizures prior to entering the study. This is to be expected since subjects with a tendency for more frequent seizures are likely to go less time 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.

Next we assess whether proportional hazards is a suitable assumption for these data. Looking at the residual plots, the treatment seems to have a decreasing effect over time, and the residuals have mean consistently greater than zero. 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.

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)	p.coef	p.PH
trt1	0.351	1.420	0.000	0.024
ager	-0.003	0.997	0.193	0.904
sex1	-0.040	0.960	0.604	0.576
nseiz.trTRUE	0.633	1.882	0.000	0.034
period.tr	-0.044	0.957	0.046	0.130



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 it drastically reduces the estimated treatment effect.

	coef	exp(coef)	p.coef	p.PH
trt1	0.218	1.244	0.047	0.412
ager	-0.003	0.997	0.178	0.901
sex1	-0.045	0.956	0.564	0.584
nseiz.trTRUE	0.494	1.638	0.000	0.386
period.tr	-0.045	0.956	0.043	0.140
trt1:nseiz.trTRUE	0.256	1.292	0.095	0.384

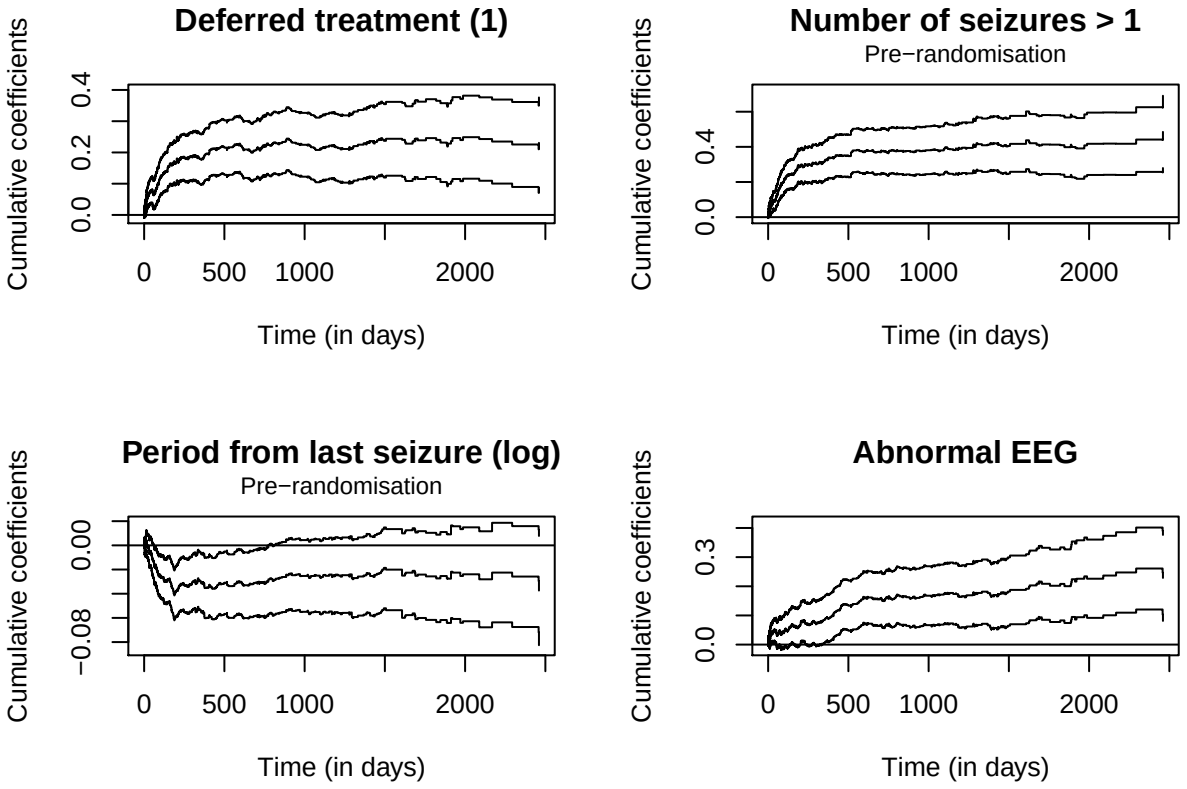
Aalen's additive model

The proportional hazards assumption underlying the 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. The model proposed by Aalen (1989) provides a flexible way to tackle these issues. In Aalen's model the effect of the covariate on the hazard function at time t is additive (rather than multiplicative as in the Cox model). In other words, the hazard function at time t is a linear function of the covariates:

$$h(t, x, \beta(t)) = \beta_0(t) + \beta_1(t)x_1 + ... + \beta_p(t)x_p.$$

Each coefficient is a function of time: as such, the model is fully nonparametric. At time t, the coefficient function describes the variation with respect to the baseline hazard function induced by a unit change in the covariates.

Aalen's additive model is widely used because of the ease of interpretation of the graphical results, which report the cumulative regression coefficient over time (with 95% confidence bands). The slope of the cumulative coefficient function indicates the effect of a variable on the outcome of interest. For what concerns the effect of deferring the treatment, the cumulative coefficient increases roughly linearly for the first two years after randomisation, then remains flat. This shows that 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 between the last seizure and the randomisation date does not have a significant effect after two years, while the effect of an abnormal EEG is roughly constant over time.



Parametric survival models

Another approach is to consider a parametric specification for the distribution of survival times. In particular, the class of accelerated failure time (AFT) models provides a natural extension to the Cox model in the case where the proportional hazards assumption does not hold. Following the introduction provided by Hosmer, Lemeshow, and May (2008), the distribution of the survival time for the i-th subject is

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

where σ is a scale parameter and ϵ_i is an error term with a prespecified distribution (its choice determining the particular type of model). The effects in this model are multiplicative on the time scale, i.e. the survival time 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

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

the effect of the covariates is to accelerate or decelerate the time to the event of interest with respect to a baseline subject (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, i.e. the effect of the covariate is protective. Therefore, the sign of the coefficient is interpreted in the opposite way to the proportional hazards models.

In this report we proposed four parametric regression models: exponential, Weibull, log-normal (where ϵ and therefore $ln(t)$ are normally distributed) and loglogistic. The same set of covariates was used for each, namely `trt`, `sex`, `age`, `nseiz.tr`,`period.tr`, `abeeg`. We used AIC to select the best-fitting model, in this case the log-normal.

	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

Having established the parametric form, we include the other covariates in the model in order to assess whether they affect the results. In particular, we expand the set of covariates to include all variables relating to the type of seizures experienced (`nsp`, `nps`, `nps`, `nmyo`, `nab`, `naab`, `ntc`, `noth`) and the type of EEG abnormality (`nsab`, `gparsp`, `gparnsp`, `fparsp`, `fparnsp`). In addition, we include an interaction term between the treatment and the number of seizures pre-randomisation. For comprehensibility we report only the significant covariates here. Many of the variables included in the model (especially those referring to an EEG abnormality) seem to be insignificant.

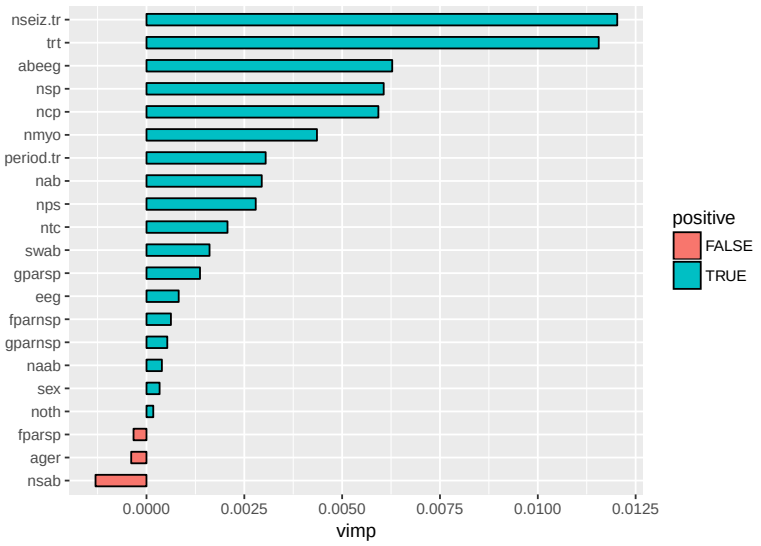
	Value	Std. Error	z	p
(Intercept)	7.096	0.491	14.441	0.000
trt1	-0.502	0.234	-2.147	0.032
nsp	-0.050	0.021	-2.406	0.016
ncp	-0.049	0.015	-3.369	0.001
nps	-0.087	0.032	-2.718	0.007
nmyo	-0.027	0.009	-3.015	0.003
nseiz.trTRUE	-0.988	0.283	-3.490	0.000
period.tr	0.141	0.054	2.621	0.009
abeeg1	-1.199	0.364	-3.294	0.001
nsab1	0.924	0.419	2.207	0.027
trt1:nseiz.trTRUE	-0.732	0.346	-2.117	0.034
Log(scale)	1.042	0.030	35.160	0.000

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 are supported by only one observation. We now include the counts of each 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 decrease after creating 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=0.002` 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. 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 were significant in both the parametric and non-parametric approaches, namely `nseiz`, `trt`, `ncp`, `nsp`, `abeeg`, and `nmyo`.

Final model

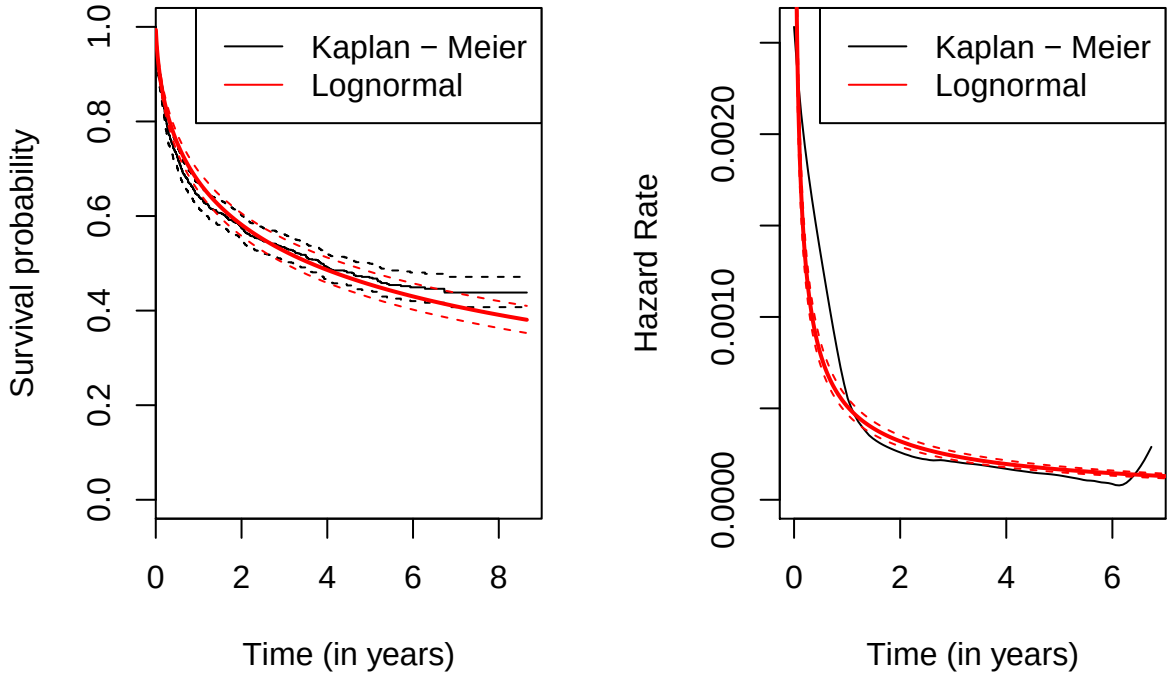
Given a parametric specification for the model and a restricted set of covariates, we fit the final model. We also check the significance of the interaction term between the treatment and the number of seizures pre-randomisation. The advantages of such a model are the parsimony and the fact that, thanks to the full parametric representation, we can easily predict the survival function for a subject given its characteristics. To aid interpretation we report the hazard ratio for each coefficient β, $HR = \exp[-\beta]$.

	Coefficient	HR	Std. Error	z	p
(Intercept)	8.286		0.202	41.07	0
trt1	-0.516	1.676	0.236	-2.183	0.029
nseiz.trTRUE	-0.708	2.03	0.253	-2.795	0.005
abeeg1	-0.596	1.815	0.176	-3.378	0.001
ncp	-0.049	1.05	0.015	-3.339	0.001
nsp	-0.046	1.047	0.021	-2.184	0.029
nmyo	-0.026	1.026	0.009	-2.895	0.004
trt1:nseiz.trTRUE	-0.633	1.882	0.348	-1.815	0.069
Log(scale)	1.06		0.03	35.79	0

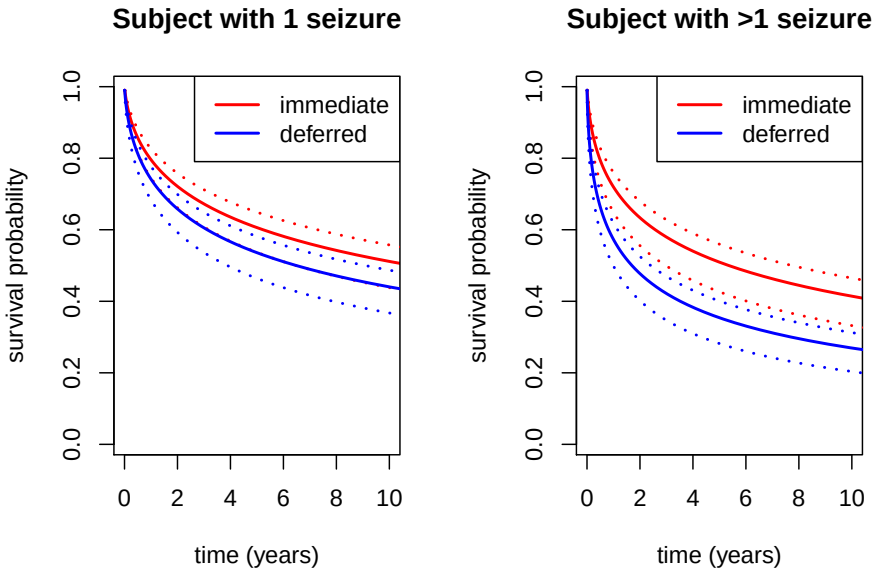
Having a deferred treatment increases the hazard of a new seizure by around 68%; the hazard for a subject with more than one seizure pre-randomisation is approximately double than the one for those with only one seizure, *ceteris paribus*. At the 0.05 significance level, the interaction term is not significant so we remove it.

The presence of an abnormal EEG induces a large change in the hazard, while the other variables concerning the type of seizures pre-randomisation give a hazard ratio close to 1 but are significant.

In this model we assume that only the location parameter of the log-normal distribution depends on the covariate and we estimate the scale parameter as 2.8850143. (We tried to estimate the scale as a function of the treatment, but it turned out to be not significant.) The hazard function for this model is non-monotonic; it exhibits a maximum depending on the scale parameter - in this case it is very close to 0 because of the magnitude of this parameter. We assess the fit graphically by comparing the survival function and hazard function estimated with our model to the Kaplan-Meier estimates. The two plots indicate that the log-normal model captures the main behaviour of the observed curves after 2 years, but it overestimates the survival probability prior to that time point.



The parametric model gives an easy way to predict the survival probability of a subject with given characteristics. For the plots below, we considered the effect of the treatment for subject with 1 or more than 1 seizures, keeping constant the values for the other variables. The immediate treatment shows a significant beneficial effect especially for those subject presenting more than 1 seizures before randomisation.



Further research

The parametric model considered seems to be a good choice for the data at hand. Nevertheless, there are some differences between the curves in the first years that may indicate the need for introducing a frailty term that takes into account the variability between subject. Another potential source of variability not addressed in the model is the `centre` variable that may account for some correlation between subjects. If we had some additonal information say about suitable 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. Even in this case we might consider to include it as a random effect.

We have considered as our outcome the time to first seizure post-randomisation. The dataset also contains information about the time to second and fifth seizure, and the time to the first tonic-clonic seizure. Additionally investigating these outcomes could enrich the results and establish the longevity of the treatment effect.

Alternatively, working with the first seizure outcome, we could transform the outcome to an indicator "seizure free for first six months Y/N" or similar, and fit a logistic model. This outcome is of interest because it determines whether or not a patient is allowed to drive again.

Another way to look at the data is to consider each patient as having their own unobserved seizure rate λ_i pre-randomisation, and consider the treatment effect in terms of how this compares to their post-treatment seizure rate, say γ_i. This is quite different to the standard medical statistics approaches we have used, but makes sense from a modelling perspective. This approach was treat in Cowling, Hutton, and Shaw (2006).

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.

Cowling, BJ, JL Hutton, and JEH Shaw. 2006. "Joint Modelling of Event Counts and Survival Times." *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 55 (1). Wiley Online Library: 31–39.

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.