Derivation and Analysis of Malaria Parasite Clearance Times

Nick Shrine

June 21, 2009

Chapter 1

Introduction

Chapter 2

Exploratory Data Analysis

2.1 Data description

The data were received as an email attachment from our contact at Glaxo Smithkline. The attachment was a SAS data set which was able to be read into R by first exporting from SAS in dbf format and then read into R using the `read.dbf` function. Table 2.1 shows the data as an R data frame.

Table 2.1: Data as an R data frame

	CENTREID	SUBJID	SEX	plantm	acttm	parct	trt	trttxt
1	001	54	M	PRE-DOSE	-2.8333	14092	A	alone
2	001	54	M	2 HOURS AFTER FIRST DOSE	2.0333	7592	A	alone
3	001	54	M	4 HOURS AFTER FIRST DOSE	4.0500	1170	A	alone
4	001	54	M	6 HOURS AFTER FIRST DOSE	6.0000	52	A	alone
5	001	54	M	8 HOURS AFTER FIRST DOSE	8.0500	0	A	alone

The columns are:

- `CENTREID` - The centre at which the study was undertaken, eithe centre '001' or '002'.
- `SUBJID` - An identifier for each subject (patient) participating in the trial. There are 43 subjects.
- `SEX` - The sex of each subject, either 'M' or 'F'.
- `plantm` - The planned time of taking a blood sample for measuring parasite load, either 'PRE-DOSE' i.e. before the drug treatment is administered, or 'n HOURS AFTER FIRST DOSE'.
- `acttm` - The actual time in hours that the blood sample was taken relative to the time the treatment was administered, hence all pre-dose times are negative.
- `parct` - The parasite count per μL of blood.
- `trt` - An indicator of which treatment was used 'A' or 'B'.
- `trttxt` - A description of treatments 'A' and 'B', namely 'alone', meaning that a single drug was used or 'combi' meaning that a combination treatment was used.

2.1.1 The Parasite Count

One of the first questions addressed was whether the parasite count is a true count, and thus Poission statistics would be applicable, or whether it is a derived measurement. Our contact informed us that the method used to arrive at the `parct` values was that a microscopist would count a chosen area

of a slide of blood, counting parasites (N_p) and white blood cells (N_w). The number of white blood cells in a μL of blood (ρ_w) is automatically counted by a machine. Accordingly the number of parasites in a μL of blood (`parct`) is given by:

$$\texttt{parct} = \frac{N_p}{N_w}\rho_w$$

and thus we cannot treat this derived measurement as a count for modelling purposes.

We were also informed that the white blood cell count is right skewed and so we might expect that the parasite count per μL will be also. Table 2.2 shows the pre-treatment parasite counts in the subjects from each test centre and of each sex. The M^* row for centre 002 is where one large outlying value of 196029 has been removed. It can be seen that for 3 cases the mean is larger than the median meaning that the distributions are right skewed. This is to be expected for non-negative data such as this. When model fitting to this data we may have to choose some transformation of the parasite count such as taking logarithms.

Table 2.2: Pre-dose parasite counts

Centre	Sex	N	Mean	Median	SD	1st Qu.	3rd Qu.
001	M	14	27060	20960	17820.9	16750	24830
	F	10	29410	25170	16221.2	19700	30700
002	M	8	50540	23290	63679.9	12240	58290
	*M**	*7*	*29750*	*20610*	*26436.6*	*11180*	*38580*
	F	11	26110	27360	17262.4	11860	30400

2.2 Preliminary Analysis

2.2.1 Deriving PC90

The project specification says that the endpoint of primary importance is PC90, the time to achieve a reduction of the parasitaemia by 90% of baseline level. As a first attempt at deriving estimates of PC90 from the data, simple linear polynomial fits to the logarithm of the parasite count with time from first dose were investigated (actually log(1+parct)).

It was found that a cubic was the most suitable model if we include the data only up to the first 0 parasite count. For some patients where the parasite count drops quickly to 0. A fit that includes the subsequent run of 0s would pull the cubic fit away from the most sensible estimate of PC90. It is more suitable for the purpose of estimating PC90 to only model the drop in the count to 0 using a simple cubic model.

Figure 2.1 shows the cubic fits to the log parasite count for the 43 subjects treated. The horizontal dotted line on the plots shows the PC90 level i.e. where the parasite count has fallen to 10% of its initial value. The value of PC90 i.e. the time t at which the parasite count has fallen to 10% was found by least-squares i.e. using the `optimize` function in R to minimise:

$$[log(1 + 0.1P_0) - \beta_0 - \beta_1 t - \beta_2 t^2 - \beta_3 t^3]^2$$

where P_0 is the pre-treatment parasite count, t is the time from first dose and β_i are the fitted coefficients for the models for each of the 43 patients. Table 2.3 shows the values of PC90 derived from the cubic fit to the parasite count by centre, sex and treatment.

Table 2.3: Derived PC90 values

		Treatment	
Centre	Sex	A	B
001	M	3.8, 27.9, 34.3, 9.8, 5.0, 0.8, 28.9, 4.7	9.5, 5.3, 7.7, 23.5, 9.4, 8.1
	F	21.1, 23.2, 22.5, 47.2	4.2, 8.6, 9.4, 0.9, 8.6, 12.9
002	M	4.5, 2.1, 20.6, 29.1	19.3, 10.0, 15.6, 8.9
	F	20.1, 2.2, 24.0, 28.7, 5.0, 22.2, 23.7	15.4, 8.4, 5.9, 6.3

2.2.2 PC90 ANOVA

3-way ANOVA with interactions was performed on the PC90 data over centre, sex and treatment. The results are shown in Table 2.4. It can be seen

that the only significant effect on PC90 is the treatment used (`trt`), with perhaps some effect of the interaction between sex and treatment. This is a result we might expect. If we fit a model with treatment as the only factor we find that treatment B reduces the PC90 time compared to treatment A by 8.0 hours with a 95% confidence interval of (1.9, 14.1) hours.

Table 2.4: 3-way ANOVA with interactions for PC90

Response: PC90					
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
factor(SEX)	1	49.4	49.4	0.5218	0.47488
factor(CENTREID)	1	0.1	0.1	0.0008	0.97816
factor(trt)	1	697.5	697.5	7.3708	0.01022 *
factor(SEX):factor(CENTREID)	1	57.4	57.4	0.6062	0.44146
factor(SEX):factor(trt)	1	389.4	389.4	4.1151	0.05016 .
factor(CENTREID):factor(trt)	1	143.4	143.4	1.5150	0.22658
factor(SEX):factor(CENTREID):factor(trt)	1	49.3	49.3	0.5214	0.47505
Residuals	35	3312.0	94.6		

Diagnostic plots of the residuals of the ANOVA model are shown in Figure 2.2. There does not seem to be any significant departure from normality in the distribution of residuals which suggests that we may not need a transformation for modelling the PC90 data. However, there is perhaps some evidence of heteroskedasticity in that the variance of the residuals appears to increase with increasing fitted PC90 value, hence we may need a variance-stabalising transformation.

2.3 Work Plan

The next stages of the analysis are:

- Look in more detail at deriving the PC90 estimates. Logistic regression has been suggested in the project outline which would require non-linear regression techniques.
- Look at how to identify and deal with suspicious outliers in the parasite count data and how they influence our estimates of PC90.
- Incorporate the error in PC90 estimates into the modelling of treatment effect.

Chapter 3

Derivation of Parasite Clearance Times

3.1 Introduction

In the previous progress report submitted in July 2008 the data given to us by GlaxoSmithkline was described and an explanation of how the parasite count is calculated was given. The parasite count is our dependent variable and it was determined that it cannot be treated as a true "count" for statistical purposes (e.g. Poisson modelling) as it is a value derived from other measurements.

A method for estimating the time to reduce the parasitaemia to 90% of baseline (PC90) using a cubic fit to the data was presented. These PC90 estimates were then used in 3-way ANOVA, which showed that of the 3 factors *Centre*, *Sex* and *Treatment* that only *Treatment* affected PC90 ($P < 0.05$).

This report compares different methods of estimating PC90, such as logistic fitting and linear interpolation, and tries to evaluate which is the best one to use for routine estimation of PC90 for this type of data.

3.2 Estimating PC90 for a patient

3.2.1 Cubic regression

As described in the previous progress report a cubic model was fitted to the log-transformed parasite count *parct* with time from first dose t as the explanatory variable

$$log(1 + parct) = \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3 + \epsilon \qquad \epsilon \sim N(0, \sigma^2)$$

The data was only fitted up to the first point that parasite count fell to 0 as the subsequent 0 count data is not relevant in determining the time to clearance. PC90 was then estimated from the fit by numerical optimisation with statistical software to find the value of t corresponding to the point where the fitted model crosses a parasite count of $0.1P_0$, where P_0 is the initial pre-treatment parasite count.

3.2.2 Logistic regression

The logistic model fitted is

$$log(1 + parct) = \alpha + \frac{\lambda}{1 + e^{-\beta(t-\mu)}} + \epsilon$$

This was fitted to all the data as it can model a drop from an initial count level to a level of 0, unlike the cubic model. α is the lower asymptote which we would expect to be 0. $\alpha + \lambda$ is the upper asymptote which we would expect to be P_0. β determines the rate of reduction with time and μ is the point of inflection (maximum rate of reduction). This model was fitted using non-linear least-squares routines in *R* and *SAS*.

3.2.3 Interpolation

The data points immediately above and below $0.1P_0$ were joined with a straight line fit and then the point where this line crosses $0.1P_0$ determines t at PC90.

For loglinear interpolation the same procedure was performed only on a plot of $log(1 + parct)$ vs. t.

3.3 Comparison of results

Figures 3.1 and 3.2 compare the logistic with the loglinear interpolation methods; the horizontal line is the PC90 level. It can be seen in Figure 3.1 for Centre 1 male patients on treatment B that there is close agreement between the two methods. However in Figure 3.2 for Centre 2 female patients on treatment A where the data is more "erratic" that these simple approaches can fail in two ways.

Sometimes the non-linear fitting routine fails to converge on a solution (the plots with no logistic curve). This was found to be a problem using `nls` in *R*. However, the `NLIN` routine in *SAS* seems to achieve a reasonable solution in all cases. The second way in which this approach can fail is illustrated in the plot for patient 453 in Figure 3.2 in that the parasite count temporarily dropped below the 10% level before the actual final elimination of parasites had begun. It is clear that this approach should be modified to identify the *last* time at which the parasite count drops below 10%.

Table 3.1 shows a comparison of the PC90 estimates by the four different methods looked at so far. Looking at the table the agreement in PC90 estimates looks fairly good. If we perform 6 paired-sample t-tests between pairs of methods we find that there is no strong evidence that the cubic regression and linear interpolation methods produce different estimates, likewise for the cubic and loglinear methods and the linear and logistic methods. The other combinations of methods do show evidence of difference in PC90 estimates ($P < 0.0002$).

If we repeat the 3-way ANOVA of the previous report we find that *Treatment* is the only factor that affects PC90 whichever method we use. Hence it would seem that the simplest method i.e. linear interpolation is the best method at this stage. However, as we know, the parasite count can be unreliable and depends on an operator-selected "suitable" blood sample and so just one outlying value could throw the estimate by linear interpolation out as it only uses the two values above and below the PC90 count. In this respect an estimate based on more values should be more robust.

Figure 3.1: PC90 estimates for Centre 1 male patients on treatment B

Figure 3.2: PC90 estimates for Centre 2 female patients on treatment A

Table 3.1: Comparison of PC90 estimates by 4 methods

Subject ID	Centre ID	Sex	Treatment	PC90 cubic	PC90 linear	PC90 loglinear	PC90 logistic
54	001	M	A	3.82	3.97	3.85	4.14
80	001	M	A	27.87	29.73	27.32	28.50
96	001	F	A	21.13	20.15	19.76	22.95
98	001	M	A	34.25	36.63	36.30	33.51
101	001	F	A	23.20	23.87	22.45	
140	001	M	B	9.49	10.90	9.46	10.16
150	001	F	A	22.53	23.24	21.75	23.12
162	001	M	A	9.79	9.10	8.84	
176	001	M	B	5.32	5.55	5.05	5.66
182	001	F	B	4.18	3.86	3.65	4.29
183	001	M	A	4.95	4.50	4.35	
185	001	M	B	7.70	8.33	8.09	8.13
187	001	M	A	0.83	1.93	1.53	3.05
197	001	F	B	8.64	9.87	8.15	8.35
203	001	F	B	9.44	11.27	9.69	10.30
218	001	M	B	23.54	23.76	22.77	23.92
224	001	M	A	28.86	30.02	30.01	30.26
262	001	M	B	9.40	10.74	9.40	9.85
264	001	F	B	0.88	0.96	0.85	1.43
280	001	F	B	8.61	9.92	9.04	9.72
285	001	F	A	47.24	46.76	46.52	
288	001	F	B	12.86	9.67	9.38	12.38
294	001	M	B	8.11	7.93	7.73	8.68
295	001	M	A	4.67	5.11	4.83	4.98
449	002	M	A	4.50	5.42	4.82	
453	002	F	A	20.07	22.73	21.97	23.08
462	002	F	A	2.22	2.67	2.49	
469	002	M	A	2.15	2.26	2.21	
477	002	F	A	24.03	26.10	25.08	
490	002	F	A	28.73	24.10	24.17	29.94
500	002	M	B	19.35	18.03	17.15	19.66
502	002	F	B	15.39	16.10	14.77	16.33
504	002	F	A	5.04	5.18	5.00	6.64
505	002	F	B	8.37	10.30	8.74	9.02
509	002	M	A	20.58	11.79	11.59	18.63
511	002	M	B	9.99	10.55	9.51	10.91
519	002	M	B	15.60	16.54	14.68	16.08
521	002	F	A	22.21	23.34	21.64	23.38
523	002	F	B	5.90	5.93	5.84	7.20
525	002	F	B	6.26	6.91	6.11	6.17
526	002	F	A	23.71	24.46	24.42	
530	002	M	A	29.07	29.53	28.09	29.28
532	002	M	B	8.94	8.33	8.08	9.21

3.4 Ongoing work plan

- Look at ways to pick sensible starting values for the logistic regression and check that the logistic form is valid.
- Investigate the effects of outliers on the estimation of PC90 and determine how to detect and deal with them.
- Incorporate the uncertainty in PC90 estimates into the modelling of treatment effect.
- Compare parametric and non-parametric approaches to determining which factors affect PC90.
- Look at non-parametric inference for determining confidence of PC90 estimates e.g. randomisation.

Chapter 4

Analysis of Parasite Clearance Times