

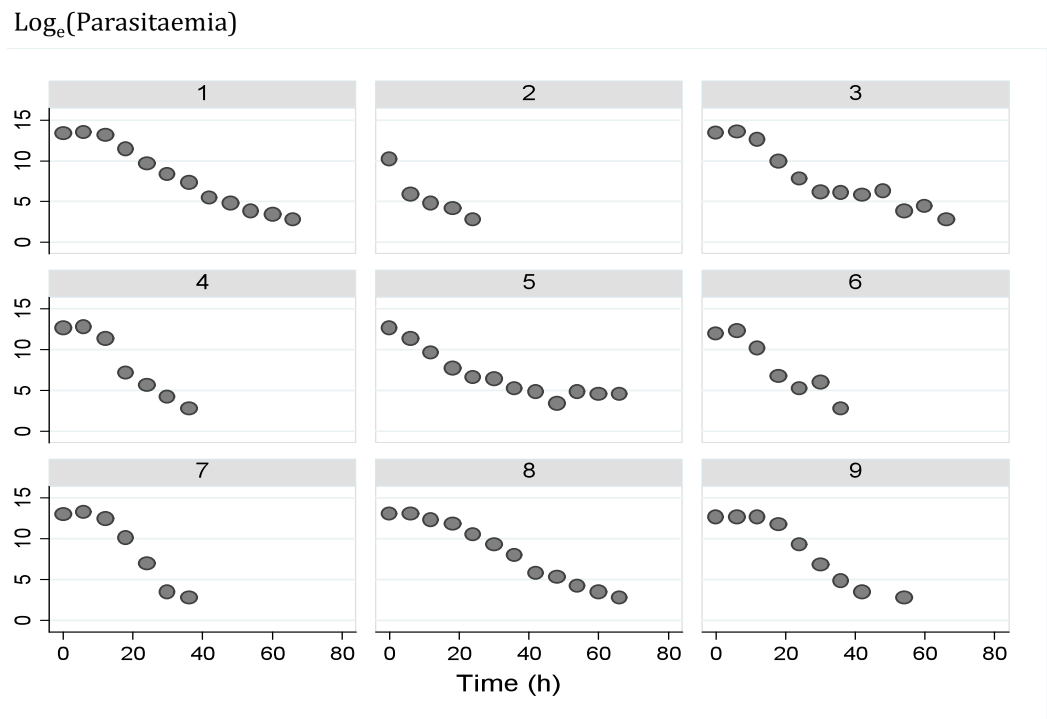
**Methodology for the WWARN Parasite Clearance Estimator**

We have developed a simple calculator that provides measures of parasite clearance from serial parasite count data. Here we present the model, terminology and methods.

**Model and Terminology**

Effective anti-malarial treatment containing artemisinin derivatives reduces parasite multiplication abruptly resulting in a rapid decline in parasite numbers. This is because of accelerated ring stage parasite clearance. Aminoquinolines have a measurable effect on ring stages, but the other antimalarials affect mainly sequestered parasites in falciparum malaria. Thus initial reductions in parasitaemia following these drugs result from sequestration. The net result is an average delay before parasitaemia falls. Thereafter parasite clearance following any effective antimalarial treatment is a first order process, resulting in killing of a fixed fraction of the parasite population in each asexual cycle, and can be considered as the reciprocal of parasite multiplication (White 1997; White, Pongtavornpinyo et al. 2009). The terminal relationship between log-transformed parasite density and time is generally linear (Day, Pham et al. 1996; Simpson, Watkins et al. 2000).

Figure 1 shows several examples of changes in natural logarithm of parasitaemia per microlitre over time, or parasite clearance profiles.



**Figure 1.** Examples of parasite clearance profiles.

Below are definitions of terms used in relation to PCE:

**Clearance rate constant** – the main part of parasitaemia clearance follows the first order process and therefore the fraction by which parasite count falls per unit time is constant. If parasitaemia at time  $t$  is given by  $P_t = P_0 \exp(-K \times t)$ , where  $P_0$  is the initial parasitaemia, then the fractional reduction is equal to  $(P_{t=1} - P_0) / P_0 = 1 - e^{-K}$ . The parameter  $K$  is called clearance rate constant and can be shown to be equal to the minus slope of the  $\log_e$  parasitaemia–time linear relationship (that is,  $K > 0$ ).

Parasitaemia declines accordingly to the first order process:

$$P_t = P_0 \exp(-K \times t)$$

After taking  $\log_e$  of each side of the equation we get:

$$\log_e(P_t) = \log_e(P_0) - K \times t.$$

**Detection limit** – for low parasite counts the thick blood smear is used and usually the number of parasites are counted against the number of white blood cells (usually 200 or 500). The detection limit obviously depends on the number of white blood cells counted. To estimate parasitaemia per microliter the following formula is used:

If  $x$  is the number of white cells counted then

Parasitaemia per microliter =

$$\text{number of parasites per slide} \times \text{white blood cell count} / x$$

Ideally the white blood cell (wbc) count is measured manually or in an automated cell counter. If this is not available then the counts are assumed to be 8,000/uL.

When the white blood cell count is assumed 8,000 for all patients, then the detection limit will be 40/uL for counting per 200 wbc and 16/uL for counting per 500 wbc.

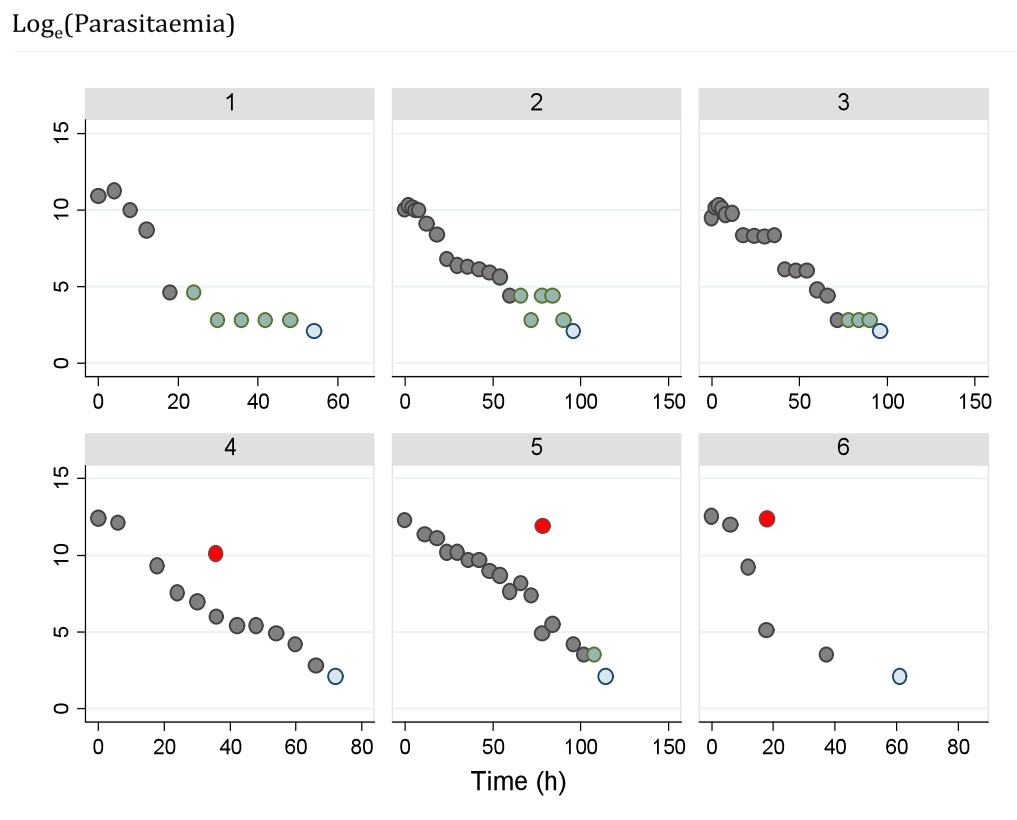
**Negative parasite slide** – when no parasites are seen while the full number of white cells have been counted then the parasite count is recorded as zero. Of course this means only that the count is below the limit of detection, although it is often reported or modeled as 0/uL.

**Outliers** – parasite counts which are not biologically possible or are highly unlikely based on other parasite measurements in the same individual. Figure 2 show examples of outlier

observations. These often result from labeling errors.

**Lag phase** – initial part of the parasite clearance profile which has a much flatter slope than the remaining part of the profile. It is important to note that a lag phase is not observed in all profiles.

**Tail** – terminal part of the parasite clearance profile when parasitaemia remains close to the detection limit (i.e. a few parasites per slide) and does not decrease over a number of measurements time-points. Tails are not observed in all profiles. For an example of a tail, see Figure 2.



**Figure 2.** Examples of parasite clearance profiles with outliers and tails: red dots represent outliers detected by program which will be excluded from analysis; green dots represent tails which will be excluded from analysis; blue dots represent observations below level of detection (measured zero parasitaemia) which are included in tobit regression analysis if there is no preceding tail identified.

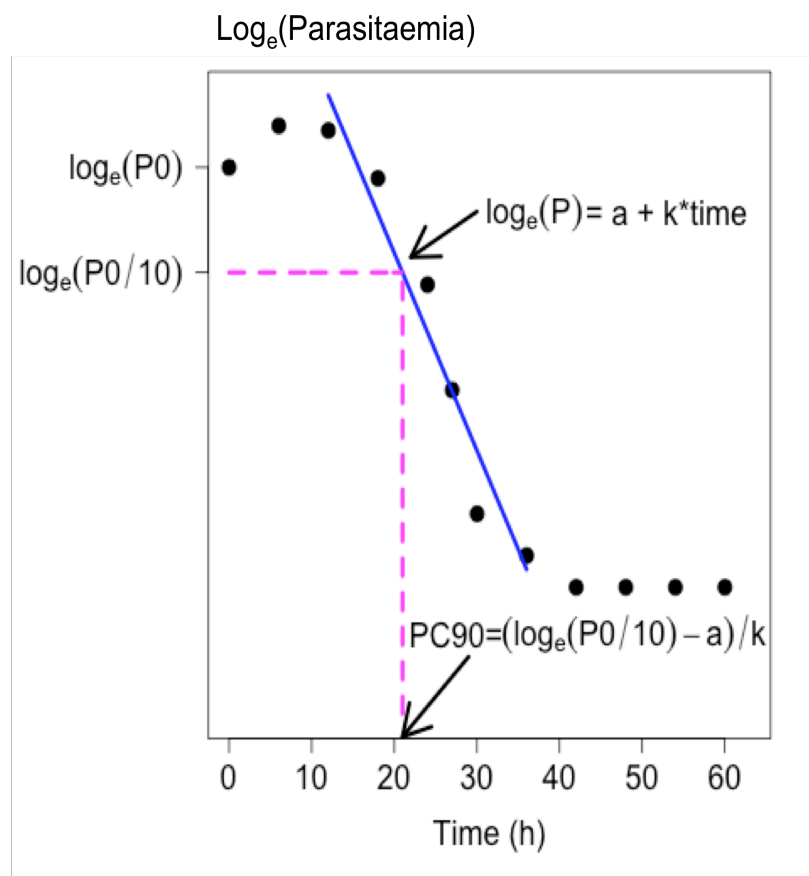
**Slope half life** – the time needed for parasitaemia to be reduced by half. This is a constant independent of the starting value of parasitaemia since reduction in parasitaemia follows the first

order process (after excluding lag phase and tail). Half life can be calculated from formula

$$T_{1/2} = \log_e(2) / K = 0.692/K$$

where K is the clearance rate constant.

**PCx** – the time it takes, based on the linear model that is fitted to the linear part of the parasite clearance profile, for the parasitaemia to be reduced to x% of the admission parasitaemia (see Figure 3). We calculate values for PC50, PC90, PC95 and PC99. Since PCx is estimated from the fitted model, in rare cases, when the linear model intercept is much lower than the measured initial parasitaemia, PC50 may not be estimable.



**Figure 3.** Schematic illustrating the calculation of the PC90 value.

### Statistical methods

When there were no zero parasitaemias recorded normal linear regression was used. When there were zero parasitaemias recorded tobit regression (Tobin 1958) was used. Only the first zero

parasitaemia which is sustained (i.e. followed by negative slides only) or the last recorded measurement was included in the analysis.

The following polynomial models of time were fitted:

1. Linear:  $b_0 + b_1 \text{time}$
2. Quadratic:  $c_0 + c_1 \text{time} + c_2 \text{time}^2$
3. Cubic:  $d_0 + d_1 \text{time} + d_2 \text{time}^2 + d_3 \text{time}^3$

If the number of observations was too few (only four positive parasite measurements) to fit cubic regression, a normal linear regression starting from the second parasite measurement was also fitted – we call this maximum regression. This was done only when the second parasitaemia measurement exceeded the first measurement (at time 0) by more than 25% since the quadratic model does not always fit well for data with a steep increase in parasitaemia.

The Akaike Information Criterion (AIC) (Akaike 1974) was used to compare models fitted to the same data. Models fitted to different numbers of observations (maximum regression and polynomial regression) were compared using the sum of squared residuals for the observations used in each model ( $\text{RSS}_{\text{shared}}$ ).

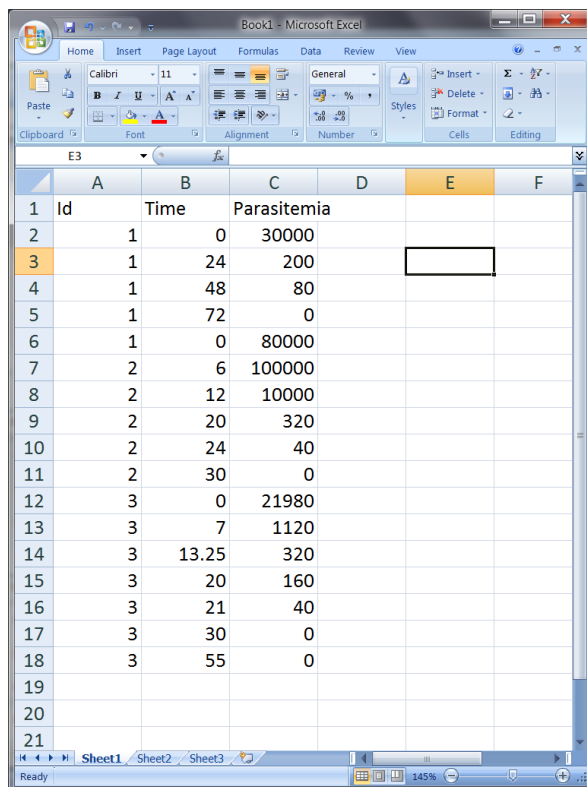
All calculations were performed in Stata and R. In order to fit tobit regression in R, the parasitaemia data needed to be transformed so that R's tobit command would converge to sensible parameter values. For each patient, parasite counts were transformed to be centered around their mean and their variance was normalised. Furthermore, starting values for the nonlinear solver were selected as the parameter estimates obtained from fitting normal regression, removing the censored value. After the model fitting was completed, the coefficients of the polynomial terms had then to be calculated from the estimates of the transformed model coefficients.

### **Input data**

The algorithm requires parasite counts per microlitre over time. At least three positive parasite counts per patient are needed for the algorithm to estimate the clearance rate constant. If, during the first 24 hours, there is a time difference between two consecutive measurements of more than 14h, the calculator will not attempt to evaluate the lag phase.

To allow appropriate handling of values below the detection limit, the detection limit for each patient is also required if zero parasitaemias are recorded. The first zero parasitaemia which is sustained or which is the last recorded observations will be included in the analysis.

Numerical data provided for the calculator need to be cleaned, to ensure there are no missing values and that counts correspond to the initial parasite clearance i.e. no parasite counts are included for the recurrent infection (if it occurs). An example of how to structure the data is shown in Figure 4. There are three columns: patient identification (anonymised) as a string variable, absolute time (in hours) since the start of treatment as a numeric variable, and parasitaemia (per microlitre) as a numeric variable. The data file should be saved in .csv format.



	A	B	C	D	E	F
1	Id	Time	Parasitemia			
2	1	0	30000			
3	1	24	200			
4	1	48	80			
5	1	72	0			
6	1	0	80000			
7	2	6	100000			
8	2	12	10000			
9	2	20	320			
10	2	24	40			
11	2	30	0			
12	3	0	21980			
13	3	7	1120			
14	3	13.25	320			
15	3	20	160			
16	3	21	40			
17	3	30	0			
18	3	55	0			
19						
20						
21						

**Figure 4.** Organization of parasitaemia data required by PCE. Data should be saved in .CSV format.

### **Data cleaning**

The entire process of data cleaning is performed in the following order:

1. Removal of data from recurrences of parasitaemia

To find data potentially associated with a recurrent infection, we remove any measurements after 7 days. Furthermore, if there is a time interval between two consecutive measurements greater than 24 hours, we remove any data after this interval. Finally, we remove the data after the last measured zero parasitaemia if the positive measurements proceeding and preceding it are separated by more than 24 hours (see Table 1).

2. Removal of trailing zeros

We identify the last non-zero parasitaemia value and then remove all of the zeros that occur after this point, accepting the very first zero.

3. Removal of tails

Tails are defined to be the terminal part of the parasite clearance profile when the parasitaemia remains close to the detection limit. We search for repeated parasitaemias below 100/uL and remove these data and any that fall between them.

4. Replacement of the first 0 value with the detection limit (DT)

We then search each patient data set for the last positive parasitaemia. If there is a zero parasitaemia that directly follows this, we replace the zero with the detection limit. This is done so that in tobit regression this zero-parasitaemia is recognized as being below level of detection.

5. Removal of extreme values

We remove extreme data points: those with times below 0 and/or values of parasitaemia outside of the range  $(0, 3 \times 10^6)$ . Values of zero parasitaemia are removed.

6. Removal of outliers

We then test for what we define as outliers - parasite measurements which are not consistent with the two immediately preceding and succeeding measurements. "Consistent" is defined by comparison of rate of change in parasitaemia to the average rate in this profile.

We calculate the normalized slope between each set of neighbouring points by taking the ratio of the rate of change between each set of two consecutive points to the average rate of change over the entire profile.

In the first 12 hours we remove a data point that is associated with the conditions

$$NormSlope_i < -20 \text{ and } NormSlope_{i+1} > 10$$

where  $NormSlope_i$  denotes the normalized slope between the  $i^{th}$  and  $(i+1)^{th}$  data points.

After the first 12 hours we relax this so that we now remove points associated with either of the following two conditions.

$$NormSlope_i < -7.5 \text{ and } NormSlope_{i+1} > 10$$

OR

$$NormSlope_i < -40 \text{ and } NormSlope_{i+1} > 3.75$$

Furthermore, at any stage during the parasite clearance we remove any point that is associated with any of the following 4 condition.

(i)  $NormSlope_i > 2 \text{ and } NormSlope_{i+1} < -10$

(ii)  $NormSlope_i > 10 \text{ and } NormSlope_{i+1} < -2$

(iii)  $NormSlope_i > 1 \text{ and } NormSlope_{i+1} < -20$

(iv)  $NormSlope_i > 50 \text{ and } NormSlope_{i+1} < 0.4$

Finally, we consider the last parasitaemia to be an outlier if the second last parasitaemia is less than 200 and the last is more than three times the second last and more than 100.

Initially these thresholds were defined based on an example data set (with approximately 100 profiles) with “noisy” parasite measurements extracted from the data of more than 3000 serial



parasitaemia –time profiles. For this data set the 4 authors decided independently on what would classify as an outlier and the values of the threshold were adjusted to meet these decisions. Subsequently the outliers detected for any datasets and parasite measurements with large residuals in linear regression model are inspected visually with the aim of adjusting the thresholds if proven necessary.

### **Algorithm**

Before model selection and fitting, the data are checked for problem data points, possible outliers, and persistent parasite tails in an automated fashion. Having determined that it is appropriate to fit a model to the dataset, we then proceed with model fitting and estimation of the clearance rate constant (K) and duration of lag phase (Tlag). We summarise our methodology in the following steps:

#### **For each patient separately:**

**Step 1:** Perform data cleaning, as described above

All further steps are performed on data with outliers, tails, extreme values and trailing zeros removed.

**Step 2:** Perform checks to see if the clearance rate constant can be estimated:

- (i) Number of non-zero parasite measurements less than 3
- (ii) Initial parasitaemia too low  
 $\text{initial parasitaemia} < 1000 \text{ parasites per microliter}$
- (iii) Final recorded parasitaemia too high and no zero parasitaemia recorded  
 $\text{final parasitaemia} \geq 1000 \text{ parasites per microliter}$
- (iv) A zero has been replaced and the last positive parasitaemia is too high and the replaced zero is uninformative (defined to be when the confidence interval of the normal linear regression fitted to all the data (excluding the replaced zero) includes the location of the zero – see Figure 5)  
 $\text{last non-zero parasitaemia} \geq 1000 \text{ parasites per microliter}$

Conditions (ii)-(iv) are implemented in the program through arbitrary constants, so they can be changed easily to different cut-offs if needed, for details see Table 1.

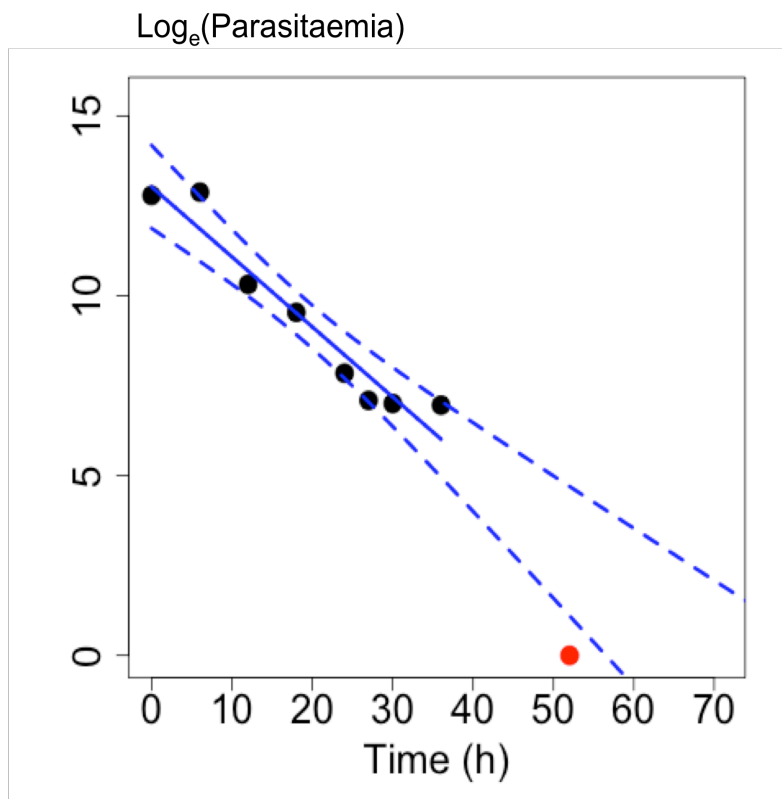


Figure 5. Schematic illustrating whether the replaced zero (shown in red) is informative. Since the red point lies outside of the 95% confidence interval (shown in blue dashed lines), the point is informative and a fit will be made.

**Step 3:** Perform additional check to see if there is enough data to estimate lag phase

- (v) There are less than three measurements in the first 24 hours or a time difference between measurements in the first 24 hours is more than 14 hours

**Step 4:** Perform model fitting

If none of the conditions (i)-(iv) in **Step 2** and (v) in **Step 3** are satisfied, fit polynomial models to the natural log-parasitaemia versus time data using Diagram 1.

If none of the conditions (i)-(iv) in **Step 2** are satisfied but condition (v) in **Step 3** is satisfied fit tobit linear regression if zero parasitaemia has been replaced by DT or linear regression otherwise.

If any of the conditions are satisfied in **Step 2** the clearance rate constant and duration of lag phase cannot be estimated. Go to **Step 6**.

**Step 5:** Estimate clearance rate constant (K) and duration of lag phase (Tlag) using Diagram 2.

**Step 6:** END

### **Estimation of Clearance Rate Constant and duration of Lag phase**

The procedure to estimate the parasite clearance rate constant depends on the shape of the parasite profile, i.e. shape of the *best* polynomial model fitted. For profiles with only three parasite counts, infrequent parasite measurements in the first 24h, profiles which are linear or exhibit concave curvature, the slope of the linear regression model or tobit linear model (where appropriate) is used as an estimate of the clearance rate constant. Profiles which exhibit convex curvature are examined further to define the lag phase. If the lag phase is not identified in the profile, then the minus slope of the linear regression model or tobit linear model is also used as an estimate of the clearance rate constant.

For profiles with convex curvature the assessment is performed on the *best* model predicted values with respect to changes in *pair-slopes* - slopes calculated between two neighbouring data points. Most *pair-slopes* are negative since the parasitaemia is decreasing and the smallest slope corresponds to the fastest parasite clearance. First, the minimum possible *pair-slope* is found for the entire profile within the time interval of positive parasite measurements. If the profile is quite linear then all *pair-slopes* would be similar to this minimum slope but if the curvature in the profile is quite pronounced, then there would be few *pair-slopes* which are near to zero (corresponding to the flat part of the curve). If this flat part of the curve is for the initial time-points then we call it the lag phase and exclude from estimation of the clearance rate constant, however the data-point which is the upper limit for the lag phase period is included in calculations (i.e. if Tlag=12h then data-point at 12h is included in the final estimation of clearance rate constant).

### **Output of results**

For each patient, a number of statistics will be given in the *Estimates.csv* and *Estimates\_short.csv* files, for details see Tables 2 and 3.

Summary statistics of the fitted profiles and estimated parameters will be given in the WWARN Parasite Clearance Estimator Report. Additionally, figures with log-parasitaemia over time, fitted

linear model used to estimate clearance rate constant and removed outliers will be presented for each patient. Removed tails are not presented in the figure since they could expand the x-axis so that the data contributing to clearance estimation would be cramped. However information about the tails is already given in the *csv* file in column *Outlier* and should be interpreted according to information provided in Table 2.

### **Program validation**

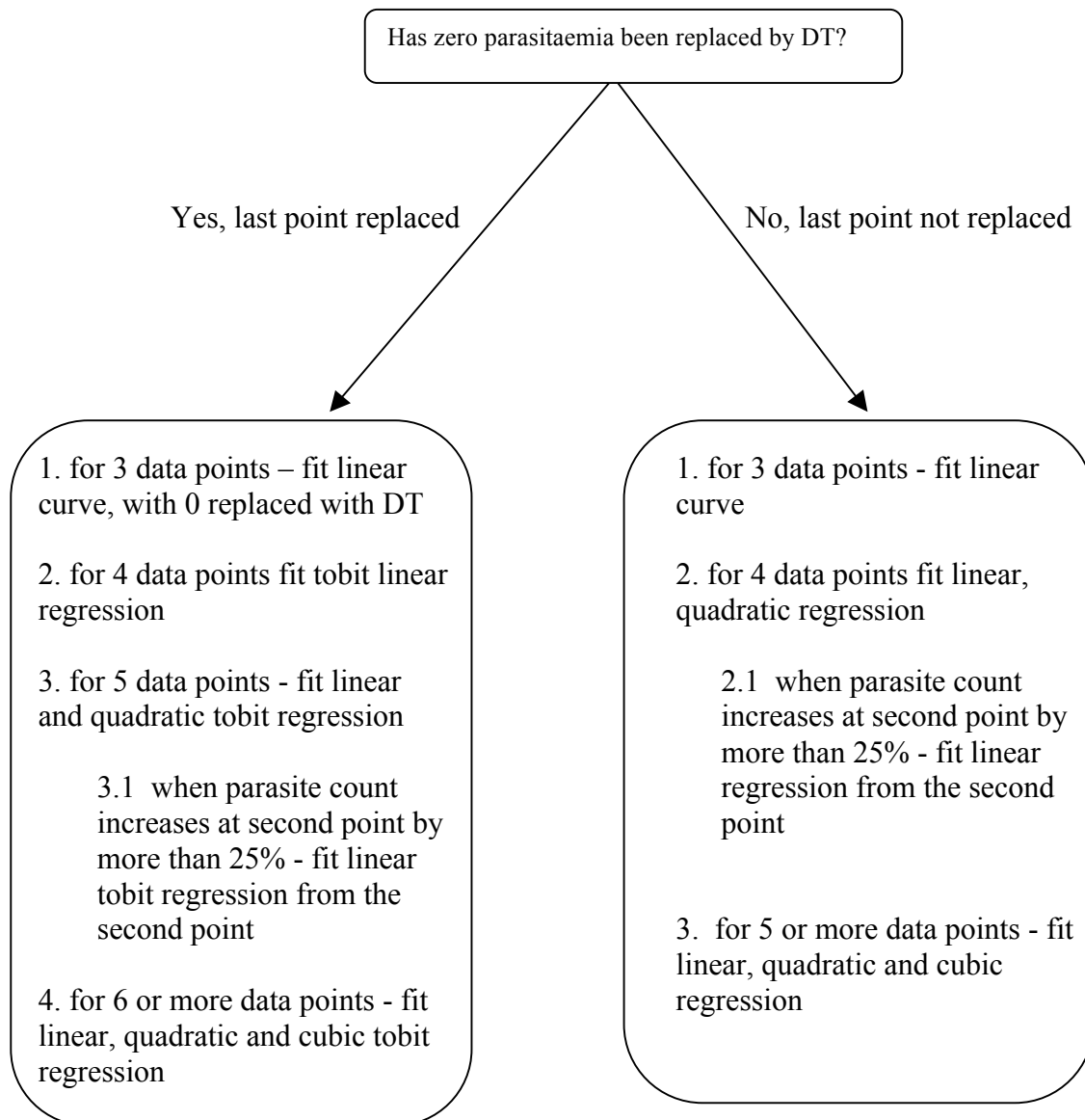
The parasite clearance calculator was programmed independently by two scientists: Dr Jennifer Flegg in R and Dr Kasia Stepniewska in Stata. Results at multiple stages of the calculations (estimates of the polynomial models, selection of the final model, estimates of the lag phase, estimates of the clearance rate) were compared for over 3000 profiles and only when all results were the same to the third decimal place, was the calculator accepted as final.

### **Acknowledgments**

We would like to thank Drs Prakaykaew Charunwatthana, Nick Day, Arjen Dondorp, Rick Fairhurst, Paul Newton, Harald Noedel, Francois Nosten and Nguyen Hoan Phu, who provided parasite data which allowed us to test our algorithm and Stata- and R- programs.

We would also like to thank the following for assistance with the R code: Dr Marcel Wolbers and the R-help forum.

**DIAGRAM 1**



---

## DIAGRAM 2

**Step 1:** Find *Best* model

*Best* model is defined as model with smallest AIC or  $RSS_{\text{shared}}$  among fitted models, where appropriate.

**Step 2:** Identify *possibly convex* models

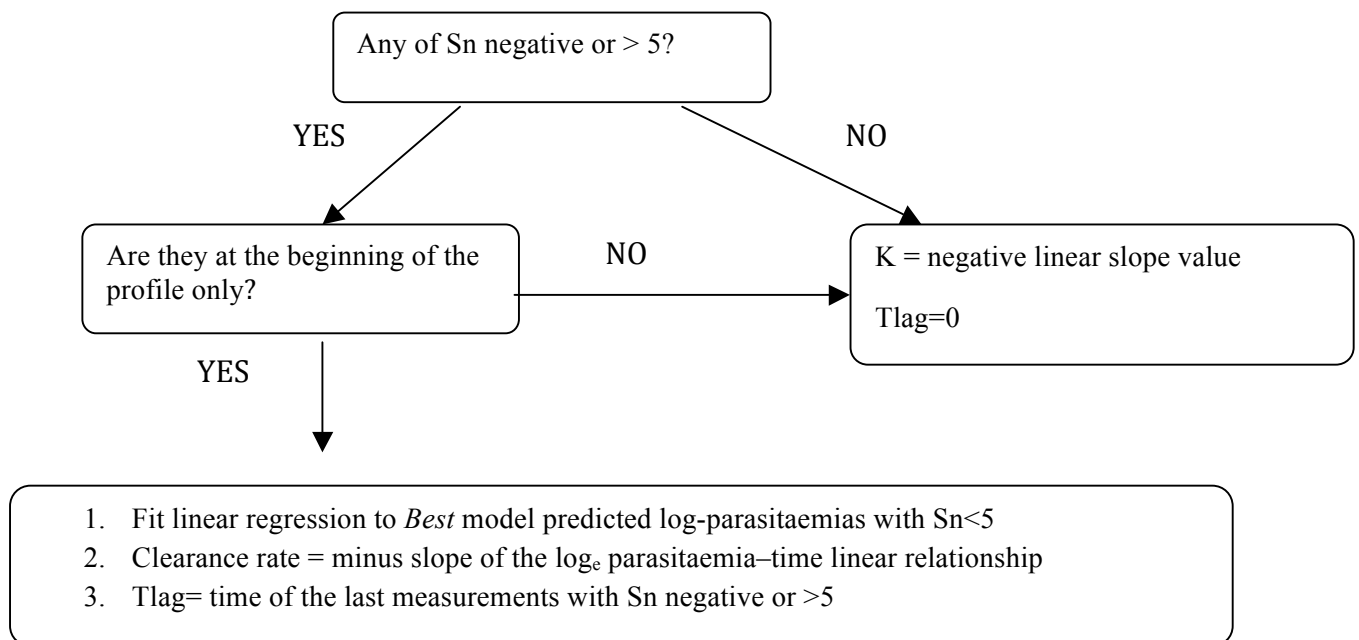
*Possibly convex* models are cubic models or quadratic models with negative concavity

**Step 3:** If model is NOT *possibly convex*:

$K$  = minus slope of the  $\log_e$  parasitaemia–time linear relationship;  $T_{\text{lag}}=0$ ; GO TO **Step 5**

**Step 4:** If model is *possibly convex*

- 4.1 For each log-parasitaemia predicted by the *Best* model  $y_i$  (but excluding any measured zero parasitaemias), calculate slope  $S_i$  between this point and the preceding predicted value
- 4.2 Find the most negative slope,  $S_{\text{max}}$
- 4.3 Calculate normalised slopes  $S_n = S / S_{\text{max}}$
- 4.4 Find clearance rate using the chart below



**Step 5:** END

---

**Table 1. Constants used**

<b>Constant</b>	<b>Value</b>	<b>Description</b>
Fact	0.25	The factor by which the second parasitaemia level must exceed the first (when there are only 4 data points) in order that a straight line model is fitted from the maximum parasite density recorded. Note $0.25 = 25\%$ increase.
Threshold2	1000	The threshold for the first parasitaemia value (parasites/uL). If it is below this threshold then no model is fitted.
Threshold3	1000	The threshold for the last parasitaemia level (parasites/uL). If it is above this threshold then no model is fitted.
FirstHours	24	The initial time period that should contain an adequate number of parasite measurements, otherwise a lag time cannot be estimated and the default linear model is fitted.
NeededInFirstHours	3	The minimum number of measurements required in FirstHours to estimate lag phase
MaxDiffInFirstHours	14	The maximum difference between measurements allowed during the FirstHours so that a lag time could be estimated.
DaysRecurrence	7	Data after DaysRecurrence days are removed as “possibly” representing an episode of recurrence.
TimeDiffRecurrence	24	The threshold for time interval between two consecutive positive parasite counts when at least one negative parasite slide was recorded between them. If time interval is greater than TimeDiffRecurrence then any parasite counts after recorded zeros are excluded from estimation.

**Table 2. Variables provided in *Estimates.csv* file**

Name	Description	Coding /Remarks
Id	Patient id	As given by investigator
Time	Time of measurement in hours	As given by investigator
Para	Measured parasitaemia (per microliter)	As given by investigator
Lpara	Log <sub>e</sub> measured parasitaemia	Calculated from values given by the investigator
Detect	Detection limit used	As given by investigator
Outlier	Outlier/Tail detection	0 = data is included (no outlier detected) 1= extreme value 2=outlier 3=tail 4=recurrence episode 5= final sequence of zeros
Estimation_ summary	Indication of whether clearance estimation was successful	0 = lag estimation successful 1 = estimation successful but no lag estimation attempted 2 = no estimation
No_Fit	Reason for not fitting a model	1= Not enough data points 2=First parasitaemia < 1000 3 = Last positive parasitaemia > 1000
Tlag	Duration of lag phase in hours	
Clearance_rate_ constant	Estimated clearance rate constant (K) (1/hours)	Clearance rate constant = - slope of the final model after exclusion of outliers, lag phase and tail.
SE_clearance	Standard error of Clearance_rate_constant	
Intercept_tlag	Intercept at time = Tlag	
Slope_half_life	Estimated time in hours it takes for the parasitaemia to decrease by half (50%)	
PC50	Estimated time in hours for parasitaemia to reach 50% of its initial value	



PC90	Estimated time in hours for parasitaemia to reach 90% of its initial value	
PC95	Estimated time in hours for parasitaemia to reach 95% of its initial value	
PC99	Estimated time in hours for parasitaemia to reach 99% of its initial value	
Predicted	Predicted log parasitaemia from the final model (excluding tail and lag phase if identified)	
Linear_slope	Slope of the linear regression or tobit linear regression model after exclusion of tails and outliers but not measurements in the lag phase.	
SE_linear_slope	Standard error of Linear_slope	
Intercept_linear	Intercept of the linear regression or tobit linear regression model after exclusion of tails and outliers but not measurements in the lag phase.	
R2_linear	R2 statistic from linear regression; measurements below detection limit are excluded	
Predicted_linear	Predicted log parasitaemia from the linear regression model or tobit linear regression model	

**Table 3. Variables provided in *Estimates\_short.csv* file**

Name	Description	Coding /Remarks
Id	Patient id	As given by investigator
Estimation summary	Indication of whether clearance estimation was successful	0 = lag estimation successful 1 = no lag estimation attempted 2 = no estimation
Excluded observations	Indication if there were any data-points excluded from estimation	0 = all points included 1 = outlier detected in data 2 = tail detected in data 3 = both tail and outlier detected in data
No Fit	Reason for not fitting a model	1= Not enough data points 2=First parasitaemia < 1000 3 = Last positive parasitaemia > 1000
Tlag	Duration of lag phase in hours	
Clearance rate constant	Estimated clearance rate constant (K) (1/hours)	Clearance rate constant = - slope of the final model after exclusion of outliers, lag phase and tail.
SE_clearance	Standard error of clearance_rate_constant	
Intercept_tlag	Intercept at time = Tlag	
Slope_half_life	Estimated time in hours it takes for the parasitaemia to decrease by half (50%)	
PC50	Estimated time in hours for parasitaemia to reach 50% of its initial value	
PC90	Estimated time in hours for parasitaemia to reach 90% of its initial value	
PC95	Estimated time in hours for parasitaemia to reach 95% of its initial value	
PC99	Estimated time in hours for parasitaemia to reach 99% of its initial value	

## References

- Akaike, H. (1974). "A new look at the statistical model identification." IEEE Transactions on Automatic Control **19**(6): 716-723.
- Day, N. P., T. D. Pham, et al. (1996). "Clearance kinetics of parasites and pigment-containing leukocytes in severe malaria." Blood **88**(12): 4694-4700.
- Simpson, J. A., E. R. Watkins, et al. (2000). "Mefloquine pharmacokinetic-pharmacodynamic models: implications for dosing and resistance." Antimicrob Agents Chemother **44**(12): 3414-3424.
- Tobin, J. (1958). "Estimation of relationships of limited dependent variables." Econometrica **26**: 24-36.
- White, N. J. (1997). "Assessment of the pharmacodynamic properties of antimalarial drugs in vivo." Antimicrob Agents Chemother **41**(7): 1413-1422.
- White, N. J., W. Pongtavornpinyo, et al. (2009). "Hyperparasitaemia and low dosing are an important source of anti-malarial drug resistance  
Artemisinin resistance in Plasmodium falciparum malaria." Malar J **8**(253): 253.