# PEER REVIEW HISTORY

## ARTICLE DETAILS

| | |
|---|---|
| **TITLE (PROVISIONAL)** | Randomised controlled trial of weekly chloroquine to re-establish normal erythron iron flux and haemoglobin recovery in post-malarial anaemia |
| **AUTHORS** | Cox, Sharon; Nweneka, Chidi; Doherty, Conor; Fulford, Anthony; Moore, Sophie; Prentice, Andrew |

## VERSION 1 - REVIEW

| | |
|---|---|
| **REVIEWER** | M. Boele van Hensbroek,<br>Paediatrcian, ID specialist<br>University of Amsterdam<br>The Netherlands<br><br>I have no competing interests. |
| **REVIEW RETURNED** | 06-Feb-2013 |

| | |
|---|---|
| **THE STUDY** | In the limitations the difference between the sample size calculation and the actual size of the study is not mentioned. |
| **RESULTS & CONCLUSIONS** | 1) One of the main problems of the study is that the study maybe under-powered. In the method section it is stated that the sample size should be 130 (2x 65), whilst they only recruit 96 patients. It can not be excluded that the fact that they do not find a difference between the intervention and placebo arm maybe due to a power issue.<br>2) The other problem is that the paper would benefit from more focus. They could consider only to present the CQ intervention study and leave out the predictor analysis and comparison of antimalarial drugs.<br>3) Finally, it is not clear how the change in haemoglobin was calculated. I am of the opinion that it is best to calculate the Delta haemoglobin for each child (day3-day30 and day3 to day 90) and compare the mean delta Hb's between the two intervention groups. It maybe that the authors has taken this approach, but this is not clear from their method section. |
| **GENERAL COMMENTS** | This is a well conducted study with a relevant and interesting research question. I have three points that the authors may want to take on board when preparing a next version.<br><br>1) One of the main problems of the study is that the study maybe under-powered. In the method section it is stated that the sample size should be 130 (2x 65), whilst they only recruit 96 patients. It can not be excluded that the fact that they do not find a difference between the intervention and placebo arm maybe due to a power issue. This point is not discussed in their discussion section.<br>2) The other problem is that the paper would benefit from more focus. They could consider only to present the CQ intervention study and leave out the predictor analysis and comparison of antimalarial |

| | |
|---|---|
| | drugs.<br>3) Finally, it is not clear how the change in hemoglobin was calculated. I am of the opinion that it is best to calculate the Delta hemoglobin for each child (day3-day30 and day3 to day 90) and compare the mean delta Hb's between the two intervention groups. This is better then taking the mean Hb's at day 30 and 90 for each study group. It maybe that the authors has taken this approach, but this is not clear from their method section. |

| | |
|---|---|
| **REVIEWER** | Dr Brian Faragher<br>Senior Lecturer in Medical Statistics<br>Clinical Group, Liverpool School of Tropical Medicine<br><br>I have no conflict of interest that would influence my review of this paper. |
| **REVIEW RETURNED** | 25-Feb-2013 |

| | |
|---|---|
| **RESULTS & CONCLUSIONS** | There is a possible flaw in the statistical analysis of the factors potentially influencing the primary outcome - and the effect size for the primary outcome is not reported with its confidence interval. There is some doubt, therefore, as to whether this is a true negative finding or a statistical type II error. |
| **GENERAL COMMENTS** | Introduction<br><br>An excellent rationale is presented for this study and there is a clear statement of the hypothesis under-pinning the study; an explicit statement of the specific study objectives is provided later in the paper.<br><br>Study design and population<br><br>The study was conducted over two malaria transmission seasons, between which unexpected changes were made to the national malaria treatment guidelines in The Gambia. The factorial element of the original RCT design had to be abandoned in favour of a simpler two-arm design. As a consequence, children recruited in the second season did not receive a pre-randomisation therapeutic dose of either CQ-SP or ACT. A marked reduction was also observed in the malaria transmission rate in the second season, which appeared to affect baseline parasite density and haemoglobin levels.<br><br>As children recruited in both seasons were randomised on day 3, the effects of both of these events should have been equally distributed between the chloroquine (CQ) and placebo treatment arms, so technically the primary objective of the study could still be addressed – but the appropriately use the term “proof of concept” to describe their study, as both events do appear to compromise the reported findings to some extent.<br><br>Inclusions and exclusion criteria<br><br>These are fully described and appear to be appropriate for the study objectives. However, the primary statistical analysis is not based on ITT (“intention-to-treat”) principles due to the contentious decision to exclude several groups of children. This runs contrary to standard practice for the reporting of RCTs; the addition of an ITT analysis would considerably strengthen the paper. |

Field and clinical procedures

The procedures used to collect the study data are well described, although no reason is given for omitting the day 70 assessment in the second season.

Randomisation and blinding

The randomisation was appropriately generated by a person independent of the study team – but the process used was more complex than necessary and may have inadvertently created the possibility of some selection bias. The randomisation was blocked to ensure that children were equally distributed between both study arms – but as all blocks were of size 8 (a fact emphasised by the use of the letters A through H inclusive on the study packaging) this may have allowed members of the study team to try to guess the identity of the treatment in the H (and possibly even the G) envelopes. Such bias would be small and possibly only at a sub-conscious level, but could easily have been avoided completely by using blocks of differing lengths (determined randomly) and having packaging labelled with study numbers only.

Interventions

The study treatments are fully described. Double blinding was achieved successfully by using matching syrup formulations for CQ and placebo.

Primary and secondary outcomes

The primary trial outcome is stated explicitly. The secondary objectives included a comparison of the children allocated to placebo in the first malaria season, to determine whether the effects of pre-randomisation dosing with CQ-SP or ACT had the same impact on the primary outcome. There is a strong argument for also including the children allocated to placebo in the second season, to determine whether the effects of pre-randomisation dosing with CQ-SP or ACT had the same effect as no pre-randomisation treatment at all.

Sample size

A valid and correctly calculated justification is provided for the sample size used, based on the detectable effect size that the authors considered likely to be clinically significant.

Statistical analysis

The statistical methods used were valid for the study design and appropriate for the data collected. On a slightly pedantic note, 95% confidence intervals were used as well as standard deviations to summarise Normally distributed continuous measures, and it is not clear why two different logarithmic bases were used.

The decision to pool the data from all assessments between days 30 and 90 should have been based on something more formal than a simple visual inspection of the haemoglobin changes over this period. As a minimum, a regression analysis comparing time trends in the CQ and placebo arms should have been carried out. Were the pooled haemoglobin levels based on the same number of observations (replicates) for all children? If not, was any weighting used to adjust for unequal numbers of replicates? Were the day 30 values included or excluded from this pooling?

### Results

The excellent CONSORT diagram (Figure 1) elegantly highlights the effects of the enforced design changes.

The characteristics of the potentially eligible children identified in the two seasons are summarised well in Table 2. Correctly, the two cohorts were not subjected to formal statistical comparisons. This Table does indicate that, with the possible exception of malaria prevalence, the two cohorts were very similar – and the text further reveals that a similar proportion of the children with a positive malaria rapid test were randomised to CQ or placebo (65/101 = 64% in season 1 and 31/49 = 63%).

The baseline characteristics of all children randomised to either CQ or placebo are well summarised in Table 3. Again, the two cohorts were correctly not subjected to formal statistical comparisons. The numerical differences between the groups were relatively small, with the possible exceptions of mean parasite density and prevalence of iron deficiency.

Mean changes in haemoglobin levels are shown for the treatment groups separately (Figure 2). As the values labelled “day 90” are presumably the pooled values for days 30 to 90 inclusive, the labelling should reflect this.

Appropriate covariate adjustment was made for important a priori factors. However, using the baseline (day 3) haemoglobin levels as a covariate when the dependent variable is haemoglobin change may have created a phenomenon known as mathematical coupling. This occurs when the outcome measure shares a common element with one or more covariates and can produce totally fallacious results; in this case, baseline haemoglobin is a covariate and is also used in the calculation of the outcome variable. The other covariates included in the models may have reduced the impact of mathematical coupling in this instance, but even so baseline haemoglobin should be removed from the covariate list. The analyses done to determine the predictors of actual haemoglobin levels at day 90 do not have this problem (Tables 4a and 4b); it is legitimate to include baseline haemoglobin levels as a covariate in this situation (paradoxically, the analysis this produces correctly determines the factors associated with change in haemoglobin levels between baseline and day 90).

When important group differences are statistically non-significant, reporting the mean values for the two groups separately is insufficient; the effect size (i.e. the difference between the two groups) should be reported along with its 95% confidence interval. True negative findings are difficult to establish as there is always the possibility that a statistical type II error has occurred (i.e. there was a clinically significant difference but the study was under-powered to detect it). Confirmation is needed that the 95% confidence interval for the effect size did not include the clinically significant difference used to inform the sample size / statistical power calculation. The current regression coefficient for “treatment group” in Table 4a appears to indicate that, after adjustment for important covariates,

| | |
|---|---|
| | the effect size confidence interval does support the conclusion this a real negative finding - but this needs to be stated explicitly (with the relevant statistical information).<br><br>The effects of pre-randomisation treatment with either CQ-SP or ACT are excellently summarised in Figure 3. As the sample sizes are very small, the confidence intervals are very wide, so there may be a high risk of a type II error (i.e. while none of the differences are statistically significant, could any be clinically significant?).<br><br>Excellent graphical summaries are also provided to show the changes in the percentages of children who were qPCR positive (Figures 4a and 4b) and bone marrow changes (Figure 5). The differences between the groups shown in all of these Figures are small and probably not statistically significant, but the inclusion of (95%) confidence intervals would confirm this.<br><br>To help the unwary reader, the legends to Figures 6a and 6b should state clearly that these are box-and-whisker plots. While this format is perfectly valid, it is not clear why it is preferred here to the more conventional “means with their 95% confidence intervals” format used for all other variables.<br><br>Discussion<br><br>A well thought out Discussion is provided - but the absence of any consideration of whether this is a real negative finding or potentially a statistical type II error is a major omission. This would be the ideal place to state the effect size and its 95% confidence interval for the primary outcome, with an accompanying reflection on the implications of the end-points of the confidence interval. |

## VERSION 1 – AUTHOR RESPONSE

Dr Boele van Hensbroek
1. Limitations – difference between the sample size calculation and the actual size of the study is not mentioned.
o The 2nd limitation bullet point has been edited to make this more explicit – p6
2. Study under-powered
o We acknowledge this point. As stated above we have edited the limitations section and the discussion on p19 but please also see responses to B10 on this point.
3. The paper could be more focussed – suggest consider only presenting only the CQ intervention study and leave out the predictor analysis and comparison of anti-malarial treatment used.
o We would prefer to leave these results in the current manuscript. The comparison of the effect of the anti-malarial treatment was part of the original study design and can’t realistically be reported elsewhere. However, we have edited the manuscript to present these analyses as supplementary material p17. We would prefer to leave in the predictor analysis as this provides important context to the discussion of the main results.
4. Calculation of Delta Haemoglobin.
o In the initial analysis and for the purpose of presentation of the raw data in Figure 3 (p16 – “effect of weekly CQ on haemoglobin change during follow-up”). Delta Hb was calculated as suggested by the reviewer and simple t-tests used to test for a significant difference by treatment group. However, for further multivariable analyses of the effect of the treatment group, a random effects model was employed in which Hb at Day3, Day30 and Day90 was modelled, in which Hb at Day30 & Day90 was in effect adjusted for that at baseline (day3). Thus making maximal use of the available data to determine if there was an effect of the intervention. Adjusting for baseline levels is considered to be the optimal statistical approach for randomised studies (but not necessarily for observational studies) and is unbiased*. The text has been edited on to try to make this process more transparent to the reader. Please also see response to comment B9.

* It is mathematically true that controlling delta Hb or final Hb for baseline give essentially the same model – only difference is that the nuisance parameter for the baseline differs by one. The question as to whether we should control for baseline at all is argued. There are three models to consider: (i) ignoring baseline altogether; (ii) subtracting but not controlling for baseline and (iii) controlling for baseline. The motivation is to remove variability due to fixed differences between individuals and hence improve precision/reduce the SE. Clearly (i) fails to do this. (ii) overdoes it because it essentially doubles the variance due the measurement error; it removes the between-individual variation but adds in the measurement error in the baseline measurement. The third option controls out as much of the between-individual variation as the baseline is capable of measuring so is optimal. There can be bias issues regarding controlling for baseline in non-randomised observational studies since the baseline itself may be affected by the exposure. This in not so a randomised trial like this
[B] Dr Brian Faragher
1. Effect size of the primary outcome is not reported
o Please see response to comment 10 below.
2. Inclusions and exclusion criteria: the primary statistical analysis is not based on an "intention to treat" principle due to the contentious decision to exclude several groups of children.
o Post-randomisation 4 children during the first malaria season became ineligible and none during the 2nd malaria season. The 4 children became ineligible due to the development of a second malarial episode – in all cases before the Day 30 Hgb. We do not have Hb data at Day 30 or Day 90 on these children and therefore cannot include them in the analyses. The relatively small number and lack of observed effect did not appear to warrant attempting to model the missing data in an analysis of intention to treat. No other children were excluded post randomisation to weekly CQ or placebo. In the first malaria season there were also children who were not eligible for randomisation.
3. Field and clinical procedures: no reason is given for omitting the day 70 assessment in the 2nd season
o The day 70 assessment was not conducted in the 2nd season due to logistical restraints in relation to judged value added.
4. Randomisation & blinding: the randomisation was blocked by blocks of 8 and labelled A-H on the study packaging which could have led to potential bias by members of the study team attempting to guess the identity of the treatments H and possibly G.
o We acknowledge the reviewers comments and agree that any such bias would be small and probably at a sub-conscious level. We agree that having blocks of different sizes and the interventions only labelled with the study IDs would have avoided this. A lesson learnt.
5. The secondary objectives included a comparison of the children allocated to placebo in the first malaria season, to determine the effect of the pre-randomisation dosing with CQ-SP or ACT had the same impact on the primary outcome. There is a strong argument for also including the children allocated to placebo in the second season to determine whether the effects of pre-randomisation dosing with CQ-SP or ACT had the same effect as no pre-randomisation treatment at all.
o All children in the 2nd malaria season received treatment with ACT. We have amended the text for the statement of the secondary objective in question (p12) – that may have led to this confusion: "(4) Hb change from Day 3 to Day 90 in the placebo arm to investigate the effects of initial malaria treatment therapy". Perhaps the reviewer is suggesting that the placebo group children from Yr 2 (who all received ACT) should have been pooled with the placebo group children in Y1 that also received ACT. However, in this case we would have to adjust for year (due to the significant differences in malaria transmission and intensity between the years) and thus this extra data would not contribute to the question being asked.
6. Statistical analysis: use of 95% CIs and standard deviations are used to summarise normally distributed continuous measures and it is not clear why 2 different logarithmeic bases were used
o The use of means and SD's has been standardised in throughout the tables of the manuscript.
7. The decision to pool the data from all assessments between days 30 and 90 should have been based on something more formal than a simple visual inspection of the haemoglobin changes over this period. As a minimum, a regression analysis comparing time trends in the CQ and placebo arms should have been carried out. Were the pooled haemoglobin levels based on the same number of observations (replicates) for all children? If not, was any weighting used to adjust for unequal numbers of replicates? Were the day 30 values included or excluded from this pooling?
o The random effects model included the measurements of Hb at Day 30 and Day 90 as repeated measures, whilst adjusting for the baseline values at Day 3. Thus this model takes into account missing data. In addition, statistical proof of difference between Day 30 and Day 90 values, or a failure to show such, would not affect the validity of this model. Finally, only 8 subjects (5 in the placebo arm and 3 in the CQ arm) had missing Hb data at Day 90 and who therefore had only the

Day 30 value included in the model.

8. Mean changes in haemoglobin levels are shown for the treatment groups separately (Figure 2). As the values labelled “day 90” are presumably the pooled values for days 30 to 90 inclusive, the labelling should reflect this.

o Figure 2 shows the “raw” data and thus does not represent the “pooled data”. Please also see response A4.

9. Appropriate covariate adjustment was made for important a priori factors. However, using the baseline (day 3) haemoglobin levels as a covariate when the dependent variable is haemoglobin change may have created a phenomenon known as mathematical coupling. This occurs when the outcome measure shares a common element with one or more covariates and can produce totally fallacious results; in this case, baseline haemoglobin is a covariate and is also used in the calculation of the outcome variable. The other covariates included in the models may have reduced the impact of mathematical coupling in this instance, but even so baseline haemoglobin should be removed from the covariate list. The analyses done to determine the predictors of actual haemoglobin levels at day 90 do not have this problem (Tables 4a and 4b); it is legitimate to include baseline haemoglobin levels as a covariate in this situation (paradoxically, the analysis this produces correctly determines the factors associated with change in haemoglobin levels between baseline and day 90).

o The model that has been used for the adjusted analysis was a random effects model using the Hb values at Day 30 and Day 90 as the outcome and adjusting for the baseline value at Day 30. This is not the same as including it as a cofactor in the outcome of delta Hb. This model makes maximal use of the available data to determine if there was an effect of the intervention. Adjusting for baseline levels is standard practice for randomised studies (but not for observational studies) and is unbiased. Please see response to comment A4 above. Please also see response to comment A4 above.

10. When important group differences are statistically non-significant, reporting the mean values for the two groups separately is insufficient; the effect size (i.e. the difference between the two groups) should be reported along with its 95% confidence interval. True negative findings are difficult to establish as there is always the possibility that a statistical type II error has occurred (i.e. there was a clinically significant difference but the study was under-powered to detect it). Confirmation is needed that the 95% confidence interval for the effect size did not include the clinically significant difference used to inform the sample size / statistical power calculation. The current regression coefficient for “treatment group” in Table 4a appears to indicate that, after adjustment for important covariates, the effect size confidence interval does support the conclusion this a real negative finding - but this needs to be stated explicitly (with the relevant statistical information).

o The treatment effect size and its 95% CI of the unadjusted analysis from the two sample t-test has been added and also for the adjusted analysis of factors identified “a priori” using the random effects model within the main body of the text – p16.For the unadjusted analysis the effect size 95% CI did include the clinically significant effect size included in the sample size calculation, but the adjusted effect size did not. We have included a specific statement of this in the discussion that it is unlikely that our observation of no effect was due to a type II error resulting from inadequate power. P19

11. Excellent graphical summaries are also provided to show the changes in the percentages of children who were qPCR positive (Figures 4a and 4b) and bone marrow changes (Figure 5). The differences between the groups shown in all of these Figures are small and probably not statistically significant, but the inclusion of (95%) confidence intervals would confirm this.

o The 95% CI around the mean reticulocyte percentages over time in the different treatment groups as been added to Figure 5 and the confidence intervals around the proportions of subjects with sub-microscopic malaria parasitaemia added to Figure 4.

12. To help the unwary reader, the legends to Figures 6a and 6b should state clearly that these are box-and-whisker plots. While this format is perfectly valid, it is not clear why it is preferred here to the more conventional “means with their 95% confidence intervals” format used for all other variables.

o Box and whisker plots were selected for the neopterin data due to its non-normal distribution. The legends have been amended to reflect this.

## VERSION 2 – REVIEW

| | |
|---|---|
| **REVIEWER** | Dr E Brian Faragher<br>Senior Lecturer in Medical Statistics<br>Liverpool School of Tropical Medicine<br>Pembroke Place<br>Liverpool L3 5QA<br>UK |
| **REVIEW RETURNED** | 22-May-2013 |

| | |
|---|---|
| **GENERAL COMMENTS** | The authors have addressed the statistical points I raised in my original review. There is now no evidence of mathematical coupling, so the statistical findings are valid. The data collected at days 30 and 90 have been analysed separately instead of pooled, to give a more detailed description of the study findings. The possibility of type II errors for important statistically non-significant findings has been fully and excellently addressed in the text. I am happy to recommend this paper for publication. |