Is there sufficient scientific evidence to rule out the use of hydroxychloroquine for postexposure prophylaxis of COVID-19?

Llorenç Quintó * ^{1,2}
José Miguel Morales ³
Raquel González ^{1,2}
Clara Menéndez ^{1,2}

¹ ISGlobal, Hospital Clínic-Universitat de Barcelona, Barcelona, Spain

² Manhica Health Research Institute (CISM), Mozambique

³ Instituto de Investigación Biomédica de Málaga (IBIMA), Universidad de Málaga, Spain

* Corresponding author: Llorenç Quintó

llorenc.quinto@isglobal.org

ISGlobal, Hospital Clínic-Universitat de Barcelona
08036 Barcelona
Spain

Keywords:

hydroxychloroquine, COVID-19, postexposure, prophylaxis, efficacy, SARS-CoV-2

Abstract

Since the beginning of the COVID-19 pandemic, the use of hydroxychloroquine (HCQ) has been surrounded by a lot of controversy, both scientific and non-scientific. This has continued with the publication of two trials of HCQ for post-exposure prophylaxis of the infection, which concluded that HCQ is not efficacious to prevent SARS-CoV-2 infection, and their results are influencing public health decisions.

We have carried out a comprehensive post-hoc analysis of the statistical power of the two trials, which shows that their power to detect an effect of HCQ in preventing COVID-19 is low, not only for their observed effect size, but also for other clinically important levels of efficacy, and therefore both studies are inconclusive.

Introduction

Two clinical trials of hydroxychloroquine (HCQ) for postexposure prophylaxis (PEP) of COVID-19 have been completed so far [1,2]. The first trial (Boulware *et al.*) has been published [1], while the second trial is available as a preprint (Mitjà *et al.*) [2]. Both trials concluded that PEP with HCQ did not prevent SARS-CoV-2 disease. Due to the enormous public health relevance of the pandemic and the urgency in obtaining evidence to inform health policy and management, the findings from these two studies are influencing recommendations on pharmacological prevention of COVID-19. However, a careful review of the published trial methods and particularly of their sample size calculations reveals important weaknesses that may compromise the validity of the findings and conclusions.

We present here an analysis on the statistical power of the above-mentioned trials that may serve as an example of the need to take with caution findings from scientific research, in order to generate solid evidence for a particular indication like the use HCQ for prophylaxis of COVID-19.

Methods

This analysis was performed with Stata software version 16.1 (Stata Corporation, College Station, TX), using its 'power two proportions' facilities for power analysis for a two-sample proportions test. Here, N , K and p refer to the number of contacts, number of clusters and proportion of contacts that meet the primary outcome at the end of the study, respectively. Subscripts 1 and 2 refer to control and intervention groups (HCQ), respectively.

Results

Table 1 shows for both studies the initial sample size calculation, their recalculation, and the final sample size with summary results from a post-hoc statistical power analysis.

Calculations made with the final sample size in Boulware *et al.* trial show that the statistical power observed in the hypothesis test corresponding to the primary outcome, was 15% (Table 1). In addition, the statistical power of the final sample size was less than 80% to detect any efficacy lower than 45% (Table 1; Figure1-a).

Figure 1-b shows the sample size needed in Boulware *et al.* trial to assess the proportion of the primary outcome between the two study groups, assuming a 14.3% prevalence of infection in the control group (the proportion observed in their sample) as a function of the expected efficacy (reduction), for a two-tailed alpha error of 0.05 and 80% statistical power. A sample size of 405 contacts per group (approximately the number of participants recruited in the trial) would have been enough if the true efficacy was at least 45%, which would have allowed to conclude with a 80% statistical power (i.e. assuming a type II error of 20%) that there is no evidence of effect of HCQ, neglecting reductions lower than 45%. With 522 or 960 contacts per group, minimum efficacies of 40% or 30% respectively could have been detected with a statistical power of 80%. A

minimum sample size of 3135 contacts per group would have been needed to detect the 17% efficacy observed in the trial with 80% statistical power.

In Mitjà *et al* trial, the initial sample calculation assumes that with an enrollment target of 95 clusters per study group, 15 participants per cluster and an intraclass correlation of 1.00, the design would yield a 90% statistical power to detect a 10% difference in the incidence of infections between the study groups, with an expected incidence of 15% in the control group. According to our calculations, the statistical power of the first estimated sample size is 63% and the result of the sample size calculation should have been 188 clusters per group. In addition, although it is not mentioned in the manuscript, according to our calculations the reestimation of the sample size should have been done assuming an intraclass correlation of 0.374 (Supplementary Material). If a value of intraclass correlation of 1.00 was used again for the recalculation of the sample size, the resulting sample size should have been 579 clusters per group (Supplementary Figure S1). Thus, under the assumptions described in the manuscript and with an intraclass correlation of 0.374, the target sample size should have been 320 clusters per group, which was achieved and even slightly exceeded (Table 1).

Our post-hoc power calculations for Mitjà *et al* trial were performed excluding the contacts that were PCR positive at baseline (n=314) leaving a total of 1042 and 958 contacts, grouped in 334 and 338 clusters in the control and intervention arms respectively, with an observed proportion of PCR-confirmed symptomatic COVID-19 of 4.3% and 3.0%, respectively (Table 1). The minimum detectable efficacy with 80% statistical power is between 51% and 79% for values of intraclass correlation of 0 and 1, respectively (Table 1 and Figure 2-a). For a value of intraclass correlation of 0.374 (likely the one used for sample size recalculation), the minimum detectable efficacy with 80% statistical power is 65%. Furthermore, to detect efficacies of 50%, 40% and 30% with the final sample size, the statistical power does not reach 80% for any value of the intraclass correlation, ranging from 35% to 78%, 23% to 56% and 15% to 33%, respectively.

Figure 2-b shows the sample size required to compare the proportion of the primary outcome assuming a 4.3% prevalence of infection in the control group (the one observed in Mitjà *et al* trial with the final sample size) as a function of the expected efficacy, for a two-tailed alpha error of 0.05 and 80% power assuming 3.5 contacts per cluster. The minimum sample required to obtain 80% statistical power with 3.5 contacts per cluster when the efficacy is 50%, 40% or 30% ranges from 302 to 1054, 501 to 1754 and 945 to 3306 clusters per group, respectively.

Discussion

The findings of these two trials need to be interpreted with caution given the deviations from the expected incidence of primary outcomes, the sample sizes and the statistical power, and in the light of the high expectation regarding HCQ effect on prevention of COVID-19.

The initial sample size calculations in both trials were based on personal perceptions, and their deviations from the real values observed once the study began were driven by

the scarcity of epidemiological information on COVID-19 at the time of trials' initiation. Unfortunately, any relevant information used for the sample size recalculation that would allow to better understand the study findings, as well as the reasons for maintaining a probably too optimistic assumption of the effect at around 50%, are lacking in both trials.

The difference in the primary outcome is not statistically significant in neither of the two trials and, therefore, there is a risk of interpretive bias when concluded that the non-significant results were due to lack of effect rather than to lack of statistical power. Although it is very difficult to agree what effect size is indeed too small to be relevant from a public health perspective, the authors should have explained how likely it was to observe a significant effect, given the sample size, and given an expected or small effect size. They should have focused on assessing if the trials were sensitive enough to detect any important difference and could still contribute to the body of knowledge if the findings showed evidence that a clinically relevant effect was absent, which is not the case. Should this post-hoc analysis had shown a low statistical power for any clinically relevant difference, they should have discussed that the trials were underpowered to detect a significant difference between the two trial arms.

In Boulware *et al.* trial, sample size recalculation was made due to the higher than expected rate of infections in the control group, although the newly estimated sample was not completed due to premature stop of the trial; the latter was justified by the lack of power observed in the third interim analysis. It is always tempting to treat post-hoc power as if it was similar to the true power. However, this is meaningless since the objective of the analysis is to guarantee that the results are statistically significant in case the difference was equal to or greater than expected, rather than to obtaining a significant p-value with the observed difference. Unless based on safety reasons, the decision to prematurely stop recruitment in a trial is never straightforward, and this might be questionable if the reason to do it is based on a lower than expected number of participants or events at that point [3]. Nobody knows what had happened if they had not stopped the trial, but our results verify that the statistical power of the final sample size for detecting efficacies lower than 45% is lower than 80% (Figure 1-a).

In Mitjà *et al.* trial, despite having recruited the desired sample size, there are other inaccuracies in the assumptions made for sample size recalculation, such as the lower than expected proportion of the primary outcome in the control group and possibly a higher intraclass correlation, all of which make it necessary to do a post-hoc power calculation analysis. The question of the intraclass correlation in a cluster randomized trial (ranging from 0 to 1) is not a trivial one, since it indicates how strongly subjects in the same cluster are similar to each other. For instance, an intraclass correlation of 0 indicates that there is no correlation of responses within a cluster and the statistical power is close to that of a design that used simple random sampling and the same total sample size of participants. On the other hand, if the intraclass correlation is near 1, the contacts within clusters are very similar to each other, and the statistical power is close to that of a design that used simple random sampling and a total sample size equal to the number of clusters [4,5]. For any given true efficacy, the statistical power of a given sample size in cluster sampling gets lower as the intraclass correlation increases and,

although the value of the observed intraclass correlation in Mitja et al manuscript is not clear, we verify that the statistical power of the final sample size for detecting efficacies lower than 51% is lower than 80%, regardless of the intraclass correlation (Figure 2-a).

Furthermore to the statistical power questions, the following are additional methodological issues that may affect interpretation of findings in both trials, and that should have been discussed in the corresponding manuscripts, such as, i) the presence of potential bias due to delay in identification of contacts and initiation of treatment (left truncation or delayed entry) [6]; ii) the justification of the statistical methodology - since a survival analysis with left-truncated data is more appropriate than the one that was used [6,7]; iii) and the lack of adjustment of the primary outcome for the viral load of index cases.

Even in the context of a pandemic like the one we are suffering, the urgency in communicating informative evidence may be paradoxically negative if the results obtained from the clinical research are not conclusive, since it may prevent the continuation of the research and the generation of valuable evidence. In addition, regarding the use of HCQ for prophylaxis of COVID-19, scientific communication has been mixed with non-scientific literature, what is making it difficult to complete ongoing clinical trials on the subject.

Informative trials are most needed to provide conclusive evidence that contribute to reduce the burden of this pandemic. Urgency in getting results should be balanced against the priority of obtaining definite and valuable scientific information. Consequently, existing published results are not sufficient to rule out HCQ as an alternative for post-exposure prophylaxis of COVID19.

References

1. Boulware DR, Pullen MF, Bangdiwala AS, Pastick KA, Lofgren SM, Okafor EC, et al. A Randomized Trial of Hydroxychloroquine as Postexposure Prophylaxis for Covid-19. *N Engl J Med*. 2020;383:517–25.
2. Mitja O, Ubals M, Corbacho M, Alemany A, Suner C, Tebe C, et al. A Cluster-Randomized Trial of Hydroxychloroquine as Prevention of Covid-19 Transmission and Disease [Internet]. *Epidemiology*; 2020 Jul. Available from: <http://medrxiv.org/lookup/doi/10.1101/2020.07.20.20157651>
3. Jitlal M, Khan I, Lee SM, Hackshaw A. Stopping clinical trials early for futility: retrospective analysis of several randomised clinical studies. *Br J Cancer*. 2012;107:910–7.
4. Campbell MK, Thomson S, Ramsay CR, MacLennan GS, Grimshaw JM. Sample size calculator for cluster randomized trials. *Comput Biol Med*. 2004;34:113–25.
5. Eldridge SM, Ukoumunne OC, Carlin JB. The Intra-Cluster Correlation Coefficient in Cluster Randomized Trials: A Review of Definitions. *Int Stat Rev*. 2009;77:378–94.
6. Yang Y, Longini IM, Halloran ME. Design and evaluation of prophylactic interventions using infectious disease incidence data from close contact groups. *J R Stat Soc Ser C Appl Stat*. 2006;55:317–30.
7. Klein JP, Moeschberger ML. *Survival analysis: techniques for censored and truncated data*. 2. ed., corr. 3. print. New York, NY: Springer; 2010.

Table 1. Initial sample size, its recalculation adapted to more realistic assumptions and final sample size with summary results from a post-hoc statistical power analysis

	<i>Boulware DR et al</i>	<i>Mitjà O et al</i>
First sample size calculation		
Assumptions	$p_1 = 10\%$ $p_2 = 5.0\%$ target power = 90%	$p_1 = 15\%$ $p_2 = 5.0\%$ cluster-size = 15 intraclass correlation = 1 target power = 90%
Sample size	1500 participants (750 per group) after accounting for a 20% incidence of attrition	95 clusters per group ^(a)
Sample size recalculation		
Assumptions	$p_1 = \text{not reported}$ ^(b) effect size = 50% reduction target power = 90%	$p_1 = 6.5\%$ $p_2 = 3.0\%$ cluster-size = 3.5 intraclass correlation = not reported ^(c) target power = 80%
Sample size	956 participants	320 clusters per group ^(c)
Actual sample size		
Sample size	821 participants $N_1 = 407$ $N_2 = 414$ $p_1 = 14.3\%$ $p_2 = 11.8\%$ observed effect size (RR) = 0.83 observed efficacy ^(d) = 17%	672 clusters $N_1 = 1042$ ^(e) ; $K_1 = 334$ $N_2 = 958$ ^(e) ; $K_2 = 338$ $p_1 = 4.3\%$ $p_2 = 3.0\%$ observed effect size (RR) = 0.70 observed efficacy ^(d) = 30%
Post-hoc analysis	minimum detectable efficacy ^(f) = 45% power for 50% efficacy = 89% power for 40% efficacy = 69% power for 30% efficacy = 43% power for observed efficacy = 15%	minimum detectable efficacy ^(f) = 51% to 79% ^(g) power for 50% efficacy = 35% to 78% ^(g) power for 40% efficacy = 23% to 56% ^(g) power for 30% efficacy = 15% to 33% ^(g) power for observed efficacy = 15% to 33% ^(g)

N_1 = number of contacts in control-group; N_2 = number of contacts in the experimental-group; K_1 = number of clusters in control-group; K_2 = number of clusters in the experimental-group; p_1 = control-group proportion; p_2 = experimental-group proportion; RR = relative risk, p_2 / p_1

(a) We have not been able to replicate this calculation based on the information available.

The statistical power with 95 clusters per group for these expected proportions and 15 contacts per cluster is 63%. The number of clusters per group to get the target power of 90% in the comparison of these expected proportions with 15 contacts per cluster should be 188.

(b) Estimated through the 2nd interim analysis.

(c) Intraclass correlation should be 0.374 to get 80% power with reported sample size.

(d) Efficacy defined as $(1 - \text{RR}) \times 100$

(e) PCR-negatives at baseline.

(f) Minimum detectable efficacy with 80% power.

(g) It depends on the intraclass correlation. Lower and upper bounds of reported range for the minimum detectable efficacy were calculated for an intraclass correlation of zero and one respectively. Lower and upper bounds of power ranges were calculated for an intraclass correlation of one and zero, respectively. Only one single value within each reported range was true for this sample, but we don't know which one as we don't know the actual intraclass correlation.

Fig. 1. Statistical power (a) and sample size (b) as a function of true efficacy for a two-samples proportion test in Boulware et al trial.

Fig. 2. Statistical power (a) and sample size (b) as a function of true efficacy according to the intraclass correlation for a two-samples proportion test in Mitjà et al trial (cluster randomized design).