So an article/website is doing the rounds describing some country-level observational analysis of the relationship between hydroxychloroquine usage in the early stages of Covid-19 and deaths from that disease per million population. It purports to show a huge positive impact – reducing the chance of death by about 80% (with a relative risk ratio of about 0.2).

My post that follows is long, so here's the spoiler – the effect is probably an artefact of some combination of failing to include the right confounders, a poor choice of response variable, choices taken by the researchers to eliminate 90% of the potential data points, and the process of classifying countries by hydroxychloroquine usage. Better statistical methods for such an observational study and small sample size return **a confidence interval for the actual risk ratio of anywhere between 0.25 and 1.0** (where 1.0 means no effect), even if we concede the categorisation of countries' usage is accurate.

Linguistic kerfuffle over what is an RCT

The website claims that the natural experiment of different countries' policy and regulatory decisions around hydroxychloroquine is actually a randomised control trial (RCT). This is clearly absurd, and an attempt to seize the prestige of RCTs for a particular position. Here's one typically (and justifiably) annoyed Twitter thread on the subject.

I imagine some of the hydroxychloroquine proponents have identified they have been losing the debate recently, because of RCTs failing to find evidence in support of hydroxycholoroquine's benefits (whereas a few weak observational studies *had* found some promising effects). So perhaps this is an attempted jujitsu move to use their critics' strength against themselves. If successful this not only lets some of the prestige of RCTs rub off on their own arguments (they have already gotten coverage on Fox News), it significantly takes the edge off future criticism based on real RCTs.

This last angle reflects exactly what happened to the analytic category of 'Fake News' – remember this term first came to prominence to categorise overwhelmingly pro-Trump malicious fictional articles in the 2016 US election (eg "Pope endorses Trump"). Now President Trump has made the term his own to refer to news he dislikes in the mainstream media, and it is nearly impossible to use in its original context. Let's hope this doesn't happen to RCTs.

That misuse of RCT terminology infuriates me and many others and has been the focus of criticism on Twitter. However, putting aside its propaganda impacts described above, from a methodological point of view it's essentially a linguistic definitional quibble. In fact, there's nothing wrong with cross country comparisons of outcomes, if we put aside what we call them, and if we use the right techniques and caveats. On an autobiographical note, wanting to better understand the strengths and limitations of World Bank cross-country regressions was the driver for me studying statistics back in my overseas aid days.

Further, I've no fondness for scientific gate-keeping; I put zero weight on the fact that this 'study' isn't peer reviewed, and might be written by someone without the 'right' qualifications. Pre-publication peer review simply doesn't work, and criticisms should be of the substance of an argument not the author.

More substantive statistical issues

So, I wanted to see if there were problems with this analysis *beyond* the misuse of RCT terminology for propaganda purposes. Here's what I thought might be going on here, with the issues I think are most likely first:

1. Researcher discretionary choices in the garden of forking paths
2. A confounding factor that correlates with both countries' hydroxychloroquine policies and their Covid-19 death rates
3. Random chance
4. A genuine effect

These aren't mutually exclusive; in fact I think it is quite likely that the first three all apply and possible that all four do. Reasons 1 and 2 are why we can't take any p values at their face value – inference is conditional on these problems not existing (this doesn't mean I am against p values, I am a big supporter – just warning to

be cautious in interpreting them). Reason 3 is basically the bad luck of every now and then one expects to find a false positive; good statistical methods can protect us from this to a degree. Reason 4 reflects my observation that at least some clear-minded experts (not just full on partisan advocates) concede it is still possible there is some very small positive impact of using HCQ in the early stages of Covid-19 or as a prophylactic, awaiting the right RCT (not this one!) to find it.

There has to be *something* going on because the claim made in the original website is simply not plausible – around an 80% drop in the probability of dying if you are in a country that uses hydroxychloroquine in the early stages of Covid-19. We know this isn't possible:

1. because such a huge effect would have shown up in the individual level trials to date even if one accepts they were flawed with regard to doses and timing as claimed by hydroxychloroquine's proponents; and
2. it simply isn't biologically plausible that a single drug could have such a huge impact, given what we know of the likely mechansim.

The sample size for this study is a very thin 19 countries. It is worth stressing up front that the sample size here is the *19* countries, not their billions of inhabitants, because the variation we have in our data is all at the country level, not individual. We have to count the countries as our sample size, not the people in them, for the same reason when we survey people we count the people as the sample size not the billions of cells that comprise them – we don't have measurements at the cell level.

Lets look at the first two of my four potential explanations in a bit of detail.

1. Research discretion in the garden of forking paths

Many of my readers will be familiar with this piece by Andrew Gelman and Eric Loken on The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time. The essence of the argument is that consciously or not, researchers take a range of discretionary decisions in their analysis process that have big impacts for inference and are assumed not to exist in standard methods and reporting.

Here are the key decisions that the researchers made here that might have an impact:

The **allocation of countries** into the "widespread", "mixed" or "limited" use of HCQ in early stages of Covid-19. This is a huge driver of the analysis and obviously of critical importance. However, the classification is done in a way that seems to leave a fair bit of room for judgement if not error. See this Twitter thread for a better-informed discussion than I could give.

Use of a **crude categorisation of HCQ usage instead of a more objective and gradated measure**. For example, how many patients get HCQ early in their treatment (rather than whether or not a media release says HCQ is available). I appreciate there may be genuine data difficulties here.

The choice of **deaths per population as the response variable rather than deaths per case**. The rationale for this is that deaths are more reliably measured; I think this is fairly generally conceded. However, it's a major limitation with the study because the claimed effect of HCQ here is on patients who receive the treatment early in the course of the disease, not in becoming unwell. It's obvious that the overhwelming driver of deaths per population is how many people get infected. The far better measure here of a treatment would be deaths per case. There are simply too many steps in the process population -> cases -> deaths that might be down to things other than HCQ.

Crucially, given the study is about the impact of a *treatment*, surely it is fair to assume that cases that are potentially exposed to the treatment are counted in the official statistics? While it is almost certainly correct that many cases are undiagnosed and reported, particularly in poorer countries, surely these are also people who were not given HCQ in the early stages.

As it happens, the relationship between deaths per case and deaths per population is fairly straightforward if noisy, as seen in the next plot. Of course, simple arithmetic shows that the relationship between the two variables in this scatter plot is effectively the cases per population. Countries such as Sweden, USA, Chile and Brazil that are above the trend are there because, while their deaths per case are unremarkable, their cases per population (and hence deaths per population) are high.

If we plug the better response variable into the classical model that is probably similar to that of the original authors, the apparent HCQ effect persists but is weaker. In trying to replicate their results, I can get an estimated point estimate of the relative risk ratio (with the model that is closest to their's) of 0.12 and confidence interval of (0.05, 0.25) – similar to the unadjusted 0.13 reported in the original website. That is when I use deaths per population as the response variable. If I change the response variable to deaths per case my new point estimate is 0.22 and confidence interval of (0.10, 0.43). This is still a very big effect (too big to be plausible), but wider, and closer to one. Remember, a relative risk of 1 is no effect. I would think any relative risk ratio below 0.9 would be startlingly low, given the scrutiny the drug has already gone through.

I am very confident that deaths per case is a better response variable than deaths per population for what is being measured here, even with reporting problems with case, and this is an important example of a researcher decision impacting on the result.

33 countries with population smaller than 1 million were excluded. This seems a mistake. A *much* better solution would have been to include them, and to weight the analysis by population (if deaths per population stays the response variable) or cases (if it changes to deaths per case, as it should).

29 countries with early adoption of masks were excluded. This seems odd. On similar logic to the discussion of why deaths per case is the better measure, it is worth noting that masks are meant to stop cases, not progression of the disease. If the better response variable of deaths per case had been chosen, the people who never get infected are dropped out of both the numerator and denominator of the response, and we could have included the sample size of countries substantially.

As an aside in the bizarre world of pandemic-as-culture war, it should be noted by any HCQ advocates that the exclusion on the website of these mask-using countries is based explicitly on the understanding that masks are more effective than HCQ at reducing deaths. So while the HCQ website/article is being widely distributed by anti-maskers they may want to bear that in mind.

51 countries with very few people aged over 80 were excluded. This doesn't make sense, given the authors are adjusting the death count for age factor.

14 countries with very little spread to date of Covid-19 were excluded. This decision is understandable, although again a better approach would have been to include these countries, use deaths per cases as response, and weight the data by number of cases.

The countries with 'mixed' use of HCQ in early stage of Covid-19 were excluded. These could have been included as simply another level of the treatment, and the sample size increased to 33. As a matter of fact, this can be done and gets the results the authors would hope for ('mixed' countries get results in deaths per population that are higher than the 'widespread' use countries, but lower than the 'limited' use countries). So this doesn't seem to be making much difference to the findings.

China is excluded. I don't know why.

As can be seen, many the above challenges follow from the decision to use deaths per population as the response variable rather than deaths per case. This cascade of decisions reduces the sample size from 176 possible countries to 19.

The other big use of researcher discretion here was in **choice of variables** to include as controls. That is the focus for my second likely explanation of the problems in the study.

Taken together, these discretionary choices have a huge impact. Exactly what I can't tell as I don't have HCQ usage data on the countries that have been omitted for what I view as unsound reasons. But it is not plausible that the impact is other than very large.

2. Confounding factors

The second thing I noticed after reading the original paper was the absence of any control for economic, political and institutional variables (the first, like everyone else, was the misuse of the 'RCT' terminology). Surely, whatever is causing different performance in number of deaths from Covid-19, diversity in these variables is more important than use of this one drug.

Here's a rather obvious chart that the researchers didn't include but perhaaps should have:

We immediately see that the reportedly limited users of hydroxychloroquine, other than Mexico, are all relatively high income countries. This also draws attention to just how much these countries have in common, other than their HCQ policies. It's not obvious why higher incomes and a common cultural background might be leading to higher death rates, but it *is* obvious that there are a bunch of variables here that aren't included in the model but should be. This chart alone is the compelling proof that "essentially random" (as one defender of this paper described it in the Twitter Wars) is *not* the same as "actually random".

For completion, here's the same chart but with case fatality rate (deaths per case) on the vertical axis rather than deaths per population. We see the relationships are a little weaker all round – between GDP and case fatality rate, and between reported HCQ usage and case fatality rate. For example, USA drops vertically and Indonesia rises, weakening both relationships, and in a way that I think is more accurate a picture answering the core research question here.

This also made me think about another obvious confounder – from my casual observation, I thought some of those poor, widespread-HCQ using countries were fairly early in their outbreaks. So I added a "how long since the outbreak got really serious" variable, based on the number of days since the country passed its first death per million. This shows some slight relationship with total deaths per million (as of course it would) and with the treatment variable, so it's another confounding variable we should control for in any serious analysis,

As it happens, the hydroxychloroquine effect is robust to controlling for these confounders. Using slightly different methods for adjusting for the confounders that the authors *had* identified (eg diabetes and hypertension prevalence, stringency of other measures, etc), and using a variety of fairly basic traditional methods, I get a relative risk ratio for the impact of "widespread" use of early hydroxychloroquine of around 0.1 or 0.2 in my various models whether or not I am controlling for GDP per capita and for the days since the country passed one death per million. I used straightforward generalized linear models for this (trying both quasibinomial and quasipoisson responses, both with logarithm link functions). I didn't think it worth while trying propensity score matching or any two step estimation process because I don't have a plausible mental model for what is behind the decision of countries to adopt the 'treatment' in this case.

See later in this post for the results of more sophisticated modelling that takes better account of uncertainty, but again the two obvious confounders that I thought of and had easy access to data for did not make a difference.

I remain pretty confident that there is a problem of some important confounding variables here, even if I (and perhaps no-one) knows exactly what they are. In a sense, this is the bottom line problem with this analysis – because there wasn't any genuine random assignment, we're never going to know what actually is driving the countries' decision to use or not hydroxychloroquine, and what other variables that usage statistic is standing as a proxy for.

3. Better modelling of chance

The original analysis involved looking at scatter plots of the relationship of a few variables like population density, stringency of non-pharmaceutical measures and life expectancy to deaths per population. In the previous section I've suggested we should be controlling for further factors like GDP per capita and how long the outbreak has been serious in each country. This is all very well, but how are we to fit a regression with so many candidate variables to a dataset with only 19 data points? Using standard methods, we would not want more than one explanatory variable in a regression with this sample size.

Putting aside the problem that the small sample size was a researcher choice (see discussion above), I think we can handle the issue by statistical inference after fitting a regression by the 'lasso'. The lasso provides a budget for total absolute size of the (scaled) coefficients in your model. It tends to force them to exactly zero if they aren't contributing materially to explaining deviance, and shrinks them towards zero even if they are. Incredibly (to me), the degrees of freedom used up by 'peeking' at all the extra variables are exactly compensated for by the shrinkage of the remaining estimates towards zero. This allows some cautious statistical inference to be done with the results. I've been reading on this recently in the truly excellent Statistical Learning with Sparsity by legends Hastie, Tibshirani, and Wainwright.

For my purposes today I didn't use the new-fangled theoretical results in chapter 6 of that book, but relied on an old favourite the bootstrap. This means I am going to resample with replacement from the 19 countries, use cross-validation to find the degree of shrinkage (the lambda parameter in `glmnet`) that results in best predictive power, and return the estimated regression coefficients from that process for each resample. Doing this enough times gets me a good, robust estimation of the distribution of the estimates that will work even with this small sample size (I think) and high number of candidate explanatory variables.

My intution for the effectiveness of the bootstrap here comes in part from suspicion at some high leverage points like Mexico. As the one lower income country with reported 'limited' HCQ usage and a high death rate, I suspected that resamples that miss out on Mexico would be likely to give quite different results to those that did.

With 19 countries

Here's the code by which I fit the lasso using the 19 countries in the original website, all their implicit explanatory variables and the two new ones added by me. The full code including creating that `all_data` data frame with a combination of scraping from the original website, downloads from Our World in Data and the World Bank is available on GitHub.

First I create a subset of the data with the candidate variables considered in the original website and the 19 complete observations. Then I define a function that will take that sort of data, scrambled as per the vector `w`, do the cross validation and return the regression coefficients. This function will be used in the bootstrap routine later.

```
#==========bootstrapped lasso============
# with just the 19 observations used in the original
d1 <- all_data %>%
  filter(hcq_usage %in% c("Widespread", "Limited") &
!is.na(urban_percentage)) %>%
  mutate(widespread_hcq = as.numeric(hcq_usage == "Widespread"),
         log_deaths_per_m = log(deaths_per_m),
         log_gdp_per_capita = log(gdp_per_capita)) %>%
  select(response_column = log_deaths_per_m,
         weight_column = pop_wdi,
         days_since_bad,
         urban_percentage:hypertension_prevalence,
         age_factor,
         log_gdp_per_capita,
         widespread_hcq)

# function for fitting a glmnet lasso model to data[w, ] - to feed to
boot()
boot_glmnet <- function(data, w, return_as_vector = TRUE){

  stopifnot(names(data)[1] == "response_column")
  stopifnot(names(data)[2] == "weight_column")

  the_data <- data[w, ]

  x <- as.matrix(the_data[ , -(1:2)]) # all the data except the first
two columns
  y <- pull(the_data, response_column)
  weights <- scale(pull(the_data, weight_column), center = FALSE)

  # use cross validation to try various values of lambda.
  # The coef(mod3) will return the coefficients from the best lambda.
  mod3 <- cv.glmnet(x, y, alpha = 1, family = "gaussian", weights =
weights)
```

```
  if(return_as_vector) {
    output <- as.numeric(coef(mod3))
  } else {
    output <- coef(mod3)
  }
  return(output)
}
```

Close readers will note that I am including my potential confounding variables (GDP per capita and length of serious outbreak) and that I am weighting each row of data by the population of the country, both of which I think are important measures.

We can test that function, outside the bootstrap, by running it just once with our unscrambled `d1` dataset, which gets us this set of coefficients:

```
> boot_glmnet(d1, 1:nrow(d1), FALSE)
14 x 1 sparse Matrix of class "dgCMatrix"
                                                 1
(Intercept)                             4.35103637
days_since_bad                          .
urban_percentage                        .
average_intervention_stringency_index   .
population_density                      .
males_per_100_females                   .
gender_factor                           .
life_expectancy                         .
diabetes_prevalence                     .
obesity_prevalence                      .
hypertension_prevalence                 .
age_factor                              .
log_gdp_per_capita                      0.08602923
widespread_hcq                         -1.29477864
```

The lasso has done its job, and only GDP per capita and the dummy variable for "widespread" use of HCQ come out as different from zero. Now we can use the bootstrap to do this whole procedure 999 times and see the distribution of results

```
booted <- boot(d, boot_glmnet, R = 999)

mean(booted$t[, 14] < 0 ) # 95% of time HCQ coefficient is less than
zero
mean(booted$t[, 13] > 0 ) # 63% of time GDP coefficient is more than
zero
mean(booted$t[, 12] < 0 ) # 24% of time age factor coefficient is less
than zero

# Confidence interval for relative risk of HCQ
boot.ci(booted, type = "perc", index = 14)
exp(boot.ci(booted, type = "perc", index = 14)$percent[4:5])
```

That gets us this result, a 95% confidence interval for the risk ratio of widespread HCQ use from 0.08 to 1.00 (where 1.00 means no effect). This is pretty intuitive to me – the 19 data points show a strong pattern which is consistent with a very strong HCQ effect. But only 19 observations! – so not surprising that the data is *also* consistent with no effect at all.

```
> exp(boot.ci(booted, type = "perc", index = 14)$percent[4:5])
[1] 0.07633073 1.00000000
```

With 177 countries

So that could easily be an end to it. However, I also used the same approach with larger datasets – 33 observations (putting the "mixed" HCQ usage countries back in) and 177 (putting all countries in, and labelling the new entrants as "unknown" HCQ usage). This gets me this interesting larger set of data:

With this much larger set of data, there is more scope for other variables such as 'days since bad' to count. In general, the longer the outbreak has been, the higher the case fatality rate. This is clearly an artefact of just where many countries are in the pandemic cycle, but is worth controlling for. Importantly, with the full dataset, we see the GDP per capita coefficient 40% of the time is negative – meaning that richer countries have lower case fatality rates than poorer. This is more what I expected than was visible with the 19 countries in the original sample (where the two variables were positively related), and reinforces my view that they can't be treated as a true random sample.

```
#--------------bootstrap lasso with 176 observations------------
# using deaths per *case* as response variable
d3 <- all_data %>%
  filter(hcq_usage %in% c("Widespread", "Mixed", "Limited",
"Unknown")) %>%
  filter(!is.na(gdp_per_capita) & deaths_per_c > 0) %>%
  mutate(widespread_hcq = as.numeric(hcq_usage == "Widespread"),
         mixed_hcq = as.numeric(hcq_usage == "Mixed"),
         unknown_hcq = as.numeric(hcq_usage == "Unknown"),
         log_deaths_per_c = log(deaths_per_c),
         log_gdp_per_capita = log(gdp_per_capita)) %>%
  select(response_column = log_deaths_per_c,
         weight_column = total_cases,
         days_since_bad,
         log_gdp_per_capita,
         unknown_hcq,
         mixed_hcq,
         widespread_hcq)

boot_glmnet(d3, 1:nrow(d3), FALSE)

booted3 <- boot(d3, boot_glmnet, R = 999)
mean(booted3$t[, 6] < 0 ) # 48% of time 'widespread' HCQ coefficient
is less than zero
mean(booted3$t[, 6] > 0 ) # 0% of time 'widespread' HCQ coefficient is
greater than zero
mean(booted3$t[, 2] > 0 ) # 73% of time 'days since bad' coefficient
is more than zero
mean(booted3$t[, 3] < 0 ) # 40% of time GDP per capita coefficient is
less than zero
mean(booted3$t[, 4] < 0 ) # 53% of time 'unknown'' coefficient is less
than zero
mean(booted3$t[, 5] < 0 ) # 38% of time 'mixed' HCQ coefficient is
less than zero
boot.ci(booted3, type = "perc", index = 6)
exp(boot.ci(booted3, type = "perc", index = 6)$percent[4:5])
```

That bottom confidence interval for the risk ratio of widespread HCQ usage is now (0.25, 1.00). I think that's the best I can do with this dataset, and feel it is controlling for everything I can. The big omission is how countries were put in the different HCQ categories in the first place, and obviously the large number of cases in the 'Unknown' HCQ usage category (ie those that I added that weren't listed in the original article).

That's all folks (and source code)

So I think that's it. These country-level data, if we take the HCQ usage categorisation in good faith, are

consistent with quite a big positive impact from HCQ but are also consistent with absolutely no impact at all. In around 50% of runs of my best model with bootstrap resamples, the ‘widespread’ use of HCQ does not feature as a variable with explanatory power for deaths per case.

Basically, I don’t think the findings stand up to scrutiny at even one percent of the emphasis with which they were put by their authors. However, I do think the exercise of looking at country-level results is legitimate. It’s just full of hazards and is vanishingly unlikely to get you strong causal conclusions. What we have here is an interesting observational study that suggests 15 countries the authors identified as having ‘widespread’ use of HCQ early in Covid-19 onset have lower deaths (whether measured per case, or per population) than have 9 countries the authors regard to have ‘limited’ use. At best this is indicative, and in particular further investigation needed of how countries were categorised, and extension needed to other countries.

But no further analysis at the country level will do more than come up with indicative ideas for proper exploration with individual data from genuine randomised control trials.

BTW I think the method here of a boostrapped lasso would also effectively have weeded out in a statistically principled fashion some of the recent weak country-level regressions that have done the rounds either on Twitter and blogs or published articles – eg that one relating cabbage consumption (or something) to Covid rates. There is also a problem for regressions of US states and counties.

> A good modestly severe test for cross-country or cross-state regressions with small sample sizes and large numbers of potential explanatory variables is that the claimed relationship appears as non-zero most of the time in a bootstrapped lasso regression; and the confidence interval from that bootstrap can be a fair indicator of the strength of the relationship.