The (statistical) power of patience

The first, obvious point I want to illustrate is that generally waiting for more cases gives higher statistical power when fixing the type 1 error. Below I simulate 10000 trials under the assumption that the true vaccine efficacy is VE=70%. The success criterion required by the FDA is that with a type I error of at most 2.5% an efficacy above 30% can be proven.

The simulation below computes the shares of trials that are successful for the three cases of an analysis after either 32, 62 or 164 cases. We assume for each case that just a single analysis takes place (no interim analyses).

```
# Helper functions to transform efficacy to theta
# and the other way round
VE.to.theta = function(VE) (1-VE)/(2-VE)
theta.to.VE = function(theta) (1-2*theta)/(1-theta)

# This function simulates a single trial
simulate.trial = function(runid=1, m.max=164, VE.true = 0.3, m.analyse
= 1:m.max) {
  theta.true = VE.to.theta(VE.true)
  is.vaccinated = ifelse(runif(m.max) >= theta.true,0,1)
  mv = cumsum(is.vaccinated)[m.analyse]
  mc = m.analyse - mv
  # Returning results as matrix is faster than as data frame
  cbind(runid=runid, m=m.analyse, mv=mv,mc=mc)
}

# Simulate 10000 trials
set.seed(1)
dat = do.call(rbind,lapply(1:10000, simulate.trial, VE.true=0.7,
m.analyse = c(32,62,164))) %>%
  as_tibble

# Parameters of used prior distribution
a0 = 0.700102; b0 = 1

# Compute posterior probabilities of VE > 30%
dat = dat %>% mutate(
  posterior.VE.above.30 = pbeta(VE.to.theta(0.3),
    shape1 = a0+mv, shape2=b0+mc, lower.tail=TRUE)
)

# Compute success share for each m
dat %>%
  group_by(m) %>%
  summarize(
    share.success = mean(posterior.VE.above.30 >= 0.975)
  )

## # A tibble: 3 x 2
```

```
##       m share.success
##
## 1    32         0.525
## 2    62         0.896
## 3   164         0.999
```

We see how the statistical power (probability to reject VE <= 30%), i.e. the success probability of the trial, increases when we wait until more cases have accrued. Analyzing (only) after 32 cases would yield a success with just 52.5% probability, while analyzing (only) after 164 cases would be successful with 99.9% probability.

So there is a natural trade-off between speed and statistical power. Hence, if cases accrue faster than expected, it seems quite natural to prefer to postpone an analysis to a higher case count.

Of course, the skipped interim analysis at 32 cases was only one of a total of the 5 planned analyses at 32,62, 92, 120 and 164 cases. But if the FDA allows to adapt the success thresholds for the later analyses correspondingly, also skipping an early interim analysis will increase total power. Yet, as is illustrated further below, skipping the 32 cases interim analysis would increase overall power only by an amount that is not too large and it is not obvious whether the FDA should indeed allow adaption of the success thresholds of the later analyses.

I guess more relevant for the decision was that a failed interim test probably would have to be announced due to SEC guidelines. But why should Biontech/Pfizer take a substantial risk of having to announce bad news if they just had to wait a few days to perform an analysis with much stronger statistical power?

How convincing would be a success with only 32 cases?

I guess a stronger concern for the FDA would be that a declared early *success* at 32 cases might have been encountered with high skepticism by the interested public arguing that 32 cases are just not enough. This could have been pouring oil on the fire of vaccine critiques and may also have reflected negatively on Biontech/Pfizer.

So if both a success and a failure at the first interim analysis would likely have created problems from a public relations point of view and the required safety milestone anyway requires to wait at least until mid November for Emergency Use Authorization, it almost seems a no-brainer to wait a few days for a more powerful first efficacy analysis.

Personally, I share the view that a small case count may be very problematic from a PR point of view. Yet, I don't fully understand in which cases a small case count is objectively statistically problematic given that we control the type I error rate.

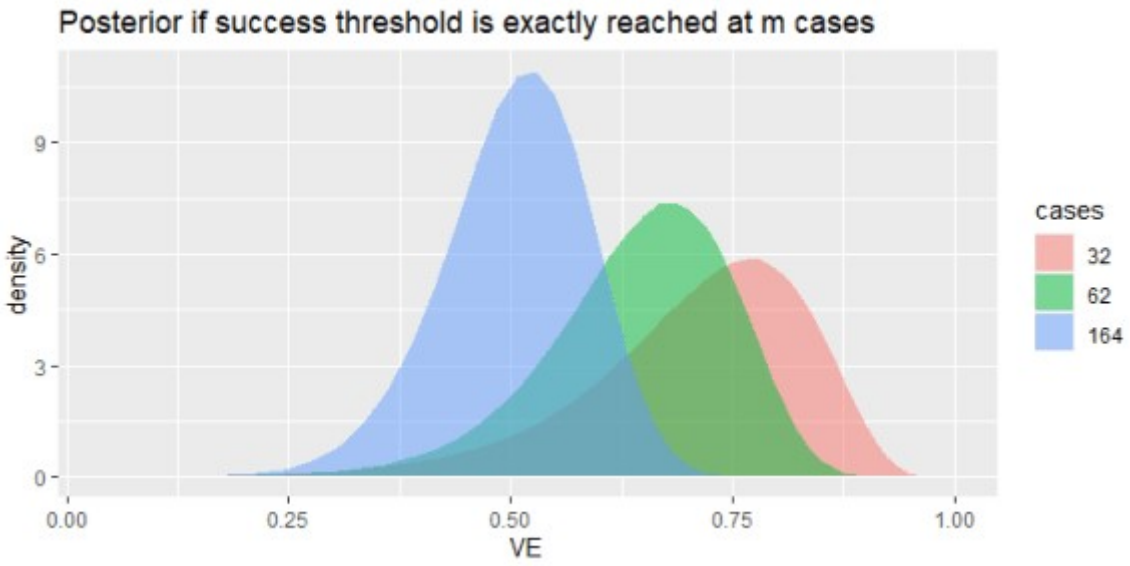
Table 5 on p. 103 in Biontech/Pfizer's study plan specifies the exact success thresholds for the 4 planned interim analysis and the final efficacy analysis. I will restrict attention to the planned analyses at 32, 62 and 164 Covid-19 cases. They would be declared a success if not more than 6, 15 or 53 confirmed Covid-19 cases, respectively, were from the vaccinated treatment group.

The following code plots the posterior distributions of the vaccine efficacy VE for each of the 3 analyses given that those success thresholds are met exactly. I assume a conservative uniform prior for the unknown parameter θ that measures the probability that a trial subject with Covid-19 was vaccinated. At the expected prior value $E \theta =0.5$ the efficacy is exactly 0.

```
a0=1; b0=1
grid = tibble(m=c(32,62,164), mv = c(6,15,53), mc=m-mv) %>%
```

```
  tidyr::expand_grid(theta = seq(0,1,by=0.01)) %>%
  mutate(
    VE = theta.to.VE(theta),
    density = dbeta(theta, shape1 = a0+mv, shape2=b0+mc),
    cases = as.factor(m)
  )

ggplot(filter(grid, VE > 0), aes(x=VE, y=density, fill=cases)) +
  geom_area(alpha=0.5,position = "identity")+
  ggtitle("Posterior if success threshold is exactly reached at m
cases")
```



We see that given that in an analysis the success threshold would be hit exactly, we should be most optimistic about the vaccine efficacy in the earliest interim analysis with just 32 cases. This finding thus does not support the skepticism against small case numbers. It rather corresponds to the observation that consistently finding significant effects with small sample sizes typically requires large effect sizes. Another reason for this finding is that Biontech/Pfizer required smaller error probabilities in the interim analysis than in the final analysis (see the previous post for details).

However, there could be problems with small sample sizes that are not accounted for in the statistical analysis. For example, in other settings a problem of small sample sizes is a larger scope for p-hacking as illustrated here. But I don't see how p-hacking could be an issue in Biontech/Pfizer's clean experimental design.

Yet, one could imagine other problems of early evaluation with small sample sizes. E.g. what if the vaccination would very commonly have side effects like fatigue or fever that make treated subjects more likely to stay at home for several days after the vaccination? They would then meet fewer people and therefore have a lower infection risk than the control group for reasons nothing to do with vaccine efficacy. While such effects should probably wash out with a longer trial duration, they may perhaps bias in a non-negligibly fashion the results if a first interim analysis takes place after very few cases.

Another problem is that the point estimate after a successful 32-cases analysis may change substantially once more observations accrue. So drafting a press release would be more complicated. Shall a high point estimate be stated, or just a quite low conservative bound? A

high estimate may sound nice, but having to reduce the efficacy as the trial proceeds may be undesirable. In theory, one could state the credible interval, but that seems to be seldom done in press releases, perhaps due to the fact that it is complicated to explain all assumptions that entered its calculations. While similar considerations also play a role for interim analysis with larger case counts, the magnitude of the uncertainties is larger, the smaller is the case count.

How much statistical power could be gained by removing the first interim trial?

We want now move towards the relegated question of how much power we could gain by dropping the first interim trial. We first simulate 100000 trials assuming a vaccine efficacy of `VE=30%`, which is the minimum efficacy that should be exceeded with a type I error of at least 2.5%.

```
# Simulate 100000 trials assuming true efficacy of only 30%
# which study plan required to exceed with a most 2.5% type I error
set.seed(1)
sim.VE30 = do.call(rbind,lapply(1:100000, simulate.trial,
  VE.true=0.3, m.analyse = c(32,62,92,120,164))) %>% as_tibble
```

We now compute the share of simulated trials that given a true vaccine efficacy of only 30% were successful, i.e. the share of trials where we would wrongly reject the null hypothesis of a 30% efficacy or less. We use the thresholds `mv_max` of the maximum number of vaccinated subjects among the `m` cases from Biontech/Pfizer's original study plan.

```
# Helper function to specify mv_max bounds.
# By default as specified in Biontech/Pfizer's original study plan.
set_mv_max = function(sim.dat,m32=6,m62=15,m92=25,m120=35,m164=53) {
  sim.dat %>% mutate(
    mv_max = case_when(
      m == 32 ~ m32,
      m == 62 ~ m62,
      m == 92 ~ m92,
      m == 120 ~m120,
      m == 164 ~ m164
    )
  )
}

# Helper function to compute shares of trials that were successful
# in interim or final analysis given mv_max bounds in sim.dat
compute.success.shares = function(sim.dat, ignore.m = NULL) {
  if (!is.null(ignore.m)) {
    sim.dat = filter(sim.dat,!m %in% ignore.m)
  }
  sim.dat %>%
    group_by(runid) %>%
    summarize(
      success = any((mv <= mv_max))
    ) %>%
    pull(success) %>%
    mean()
}
```

```
# Compute type 1 error rate given Biontech/Pfizer's specification
compute.success.shares(sim.VE30 %>% set_mv_max)
```

```
## [1] 0.0221
```

We see that that Biontech/Pfizer's specification yields a total type I error rate of 2.21%.

If we would relax the success thresholds by allowing in the final analysis at most 54 instead of only 53 Covid-19 cases to be from the treatment group, we would get the following type I error rate:

```
# Relax final analysis bound by 1 observation
compute.success.shares(sim.VE30 %>% set_mv_max(m164=54))
```

```
## [1] 0.02697
```

The 2.697% error rate would violate the required 2.5% bound.

Let us now compute the corresponding error rates assuming that we ignore the first interim analysis at 32-cases:

```
compute.success.shares(sim.VE30  %>% set_mv_max, ignore.m = 32)
```

```
## [1] 0.01745
```

```
compute.success.shares(sim.VE30 %>% set_mv_max(m164=54), ignore.m = 32)
```

```
## [1] 0.02242
```

```
compute.success.shares(sim.VE30 %>% set_mv_max(m164=55), ignore.m = 32)
```

```
## [1] 0.03015
```

We see that we could increase the threshold in the final analysis to 54 cases from the treatment group and still have an error rate of 2.242%. Increasing it to 55 cases would propel the type I error rate beyond 2.5%, however.

We now want to compare the power of the original study plan with the modified study plan that skips the first interim analysis but increases the threshold in the final analysis to at most 54 cases from the treatment group. We first compare the statistical power for a vaccine with 70% efficacy.

```
# Simulate 100000 trials given 70% efficacy
set.seed(1)
sim.VE70 = do.call(rbind,lapply(1:100000, simulate.trial,
  VE.true=0.7, m.analyse = c(32,62,92,120,164))) %>% as_tibble

# Original study plan
compute.success.shares(sim.VE70 %>% set_mv_max())
```

```
## [1] 0.99767
```

```
# Modified study plan without first interim analysis
```

```
compute.success.shares(sim.VE70 %>% set_mv_max(m164=54), ignore.m=32)
```

```
## [1] 0.99851
```

We see how dropping the first interim analysis and adjusting the final analysis threshold indeed increases the total statistical power, but just by a small amount from 99.77% in the original plan to 99.85%.

If the true efficacy were just 50% the effect would be a bit larger:

```
# Simulate 100000 trials given 50% efficacy
set.seed(1)
sim.VE50 = do.call(rbind,lapply(1:100000, simulate.trial,
  VE.true=0.5, m.analyse = c(32,62,92,120,164))) %>% as_tibble

# Original study plan
compute.success.shares(sim.VE50 %>% set_mv_max())
```

```
## [1] 0.45823
```

```
# Modified study plan without first interim analysis
compute.success.shares(sim.VE50 %>% set_mv_max(m164=54), ignore.m=32)
```

```
## [1] 0.50847
```

Now the modified analysis plan would increase the statistical power by 5 percentage points from 45.8% to 50.8%. Well, a 5 percentage point increase of the success chance for such a huge, important project seems not negligible. Ex-post it may be easy to say that assuming just 50% efficacy is unrealistically low. But how certain could one have been that the efficacy is high before analyzing the trial data?

Should the FDA allow to adapt the success thresholds if the first interim analysis is skipped?

I actually don't know whether the FDA indeed allowed to adjust the success threshold for the final analysis from 53 to 54 cases after Biontech/Pfizer agreed to skip the first interim analysis.

Should the FDA have allowed it or not? While this is ex-post not relevant given the realized tremendous efficacy (from 170 cases, only 8 were in the treatment group), I consider it an interesting academic question.

Obviously, one should not allow that a vaccine maker who already saw the unblinded data of an interim analysis, i.e. knowing how many cases were from the treatment group, can still decide whether to skip that interim analysis. While Biontech/Pfizer did not see the unblinded data when deciding to skip the first interim analysis, they knew, if I understand correctly, the development of the total case count in their experiment. While on first thought one may believe that the total case count does not reveal information about the efficacy, that is not necessarily true. The thing is that one may in principle estimate the sample efficacy when knowing the total case count from the experimental subjects and the incident rate in the total population.

For example, assume that in the analyzed period 2% of the total population got Covid-19 but only 1% from the experimental population. This may suggest a high efficacy because it is consistent with the event that only control group subjects got sick. In contrast, if also 2% of the experimental population got Covid-19, a low efficacy may be induced.

Of course, such estimates are in practice not simple. The Covid test intensity is probably higher among the experimental subjects than in the total population and the experimental subjects may substantially differ from the total population in ways that are not easily controlled for in a statistical analysis. Nevertheless, the theoretical objection remains that vaccine makers who know the total case count, also have some noisy information about the vaccine efficacy.

So I guess the cleaner approach for the FDA would be to not allow adjustment of the success thresholds after a decision to skip the first interim trial. On the other hand, given that the FDA seemed to have preferred such a skip and that also a modified treatment plan would have a type I error rate of 2.242% and thus some slack until 2.5%, allowing for such an adjustment may seem defensible.

But even without an adjustment of the later success thresholds and thus no gain in statistic power, already the reasons discussed earlier seem to make a clear case for Biontech/Pfizer to skip the 32-cases interim analysis.

Politics

It is true that a 32-case interim case analysis possibly or even likely may have taken place already before the US elections. So, not surprisingly, Donald Trump suspected a political motivation for skipping the analysis. He tweeted on Nov 10th:

> As I have long said, @Pfizer and the others would only announce a Vaccine after the Election, because they didn't have the courage to do it before. Likewise, the @US_FDA should have announced it earlier, not for political purposes, but for saving lives!

Let me conclude this post with some remarks on that tweet:

First, even if sufficient efficacy would have been announced before the election, it would, to my understanding, not have sped up vaccine deployment since the submission for emergency use authorization required anyway to wait until required safety results are available in the third week of November. So faster vaccine deployment would only be possible by forcing the FDA to reduce safety standards. But would a reduction in safety standards in expectation save lives or rather risk more deaths (possibly indirectly due to reduced willingness to get vaccinated) and risk other harmful consequences?

Second, Biontech/Pfizer did not decide whether to announce a vaccine success or not, but whether to perform the first interim analysis, without knowing its result. If the vaccine would not have been as tremendously effective as it turned out to be, the first interim analysis with 32 cases may well have been unsuccessful with high probability. For example, recent news reported an estimated average 70% efficacy for Astrazeneca's Covid-19 vaccine. If Biontech/Pfizer's vaccine would have had a true efficacy of 70%, a success would have been declared in the 32 cases interim analysis only with 37.7% probability. (You can replicate this number by adapting the code of this post using the success threshold from the original analysis plan). So from an ex-ante perspective, it is not clear in which direction an early interim analysis would have affected the election (if at all). If Trump himself believed his claims that he was the clear favorite in the election race, shouldn't he rather have preferred the absence of risk of bad news over the absence of a chance for good news?

Third, assume that it really would have been the case that a 32-cases interim analysis would have changed the presidential election result. Given that Trump was quite fond of criticizing imports from Germany and of imposing trade restrictions, it would be quite some irony of history that he lost the election because the success of a product developed in Germany was not

known before the election. (As far as I understand, the vaccine was mainly developed by Biontech, while Pfizer mainly handled the trials).

Fourth, given that it will still take substantial time until a sufficient share of the population can be vaccinated, efficient management of the Covid crisis is still highly important for several months to come. If it turns out that Joe Biden will be a better crisis manager, knowing the trial result only after the election may well have saved a substantial number of lives.