Covid vaccine: Bayesian modelling

Lecture 2 PDF version

Description

In the wake of the COVID-19 pandemic, several biotechnology and pharmaceutical companies have started collaborating on the development of a vaccine, in an unprecedented effort to deliver vital innovation in a short amount of time. The first vaccine to be approved by the FDA in December 2020 was the one developed jointly by German biotech company BioNTech and US pharmaceutical giant Pfizer. The development of the vaccine was based on a Phase I/II/III multicenter, placebo-controlled trial to evaluate the safety, tolerability, immunogenicity and efficacy of the candidate vaccine.

The Phase II/III study was designed using a Bayesian approach based on the following setup. There are two arms: one is randomised to receive two doses of a placebo, while the active arm receives two doses of the candidate vaccine. The data relevant for the efficacy analysis are the number of individuals who in each arm are confirmed COVID-19 cases, over the total number of individuals randomised to the specific arm. This can be formalised as $y_{\text{plac}} \sim \text{Bin}(\pi_{\text{plac}}, n_{\text{plac}})$ and $y_{\text{vac}} \sim \text{Bin}(\pi_{\text{vac}}, n_{\text{vac}})$, respectively. The actual measure of vaccine efficacy is defined as $\text{VE} = \left(1 - \frac{\pi_{\text{vac}}}{\pi_{\text{plac}}}\right)$, assuming a 1:1 allocation ratio in the two arms and thus $n_{\text{plac}} = n_{\text{vac}} = N$.

However, the analysis is based on a reformulation of the problem: the experimenters actually model $y = y_{\text{vac}} =$ number of COVID-19 cases among the individuals in the treatment arm over $n = (y_{\text{vac}} + y_{\text{plac}}) =$ the total number of COVID-19 cases in the two arms. This is modelled as $y \sim \text{Bin}(\theta, n)$, where

$$\begin{aligned}\theta &= \frac{\pi_{\text{vac}}}{\pi_{\text{vac}} + \pi_{\text{plac}}} \\ &= \frac{1 - \text{VE}}{2 - \text{VE}}.\end{aligned}$$

The reason for this seemingly overly-complex (and obscure) setup is that it allows the experimenters to only specify one observational model and, crucially a single prior distribution for the parameter θ — of course, once $p(\theta \mid \text{data})$ is available, using the equation above, it is straightforward (e.g. through Monte Carlo simulation) to obtain directly the posterior distribution for the main parameter of interest $p(\text{VE} \mid \text{data})$.

Sample size calculation

The determination of the sample size calculation is based on a simulation approach. The experimenter looked to determine a sample size large enough to be able to provide a probability exceeding 90% to conclude that $\text{VE} > 30\%$ with a "high probability" (we note here that the study protocol is vague on the actual threshold selected).

The simulation excercise proceeds in two steps. Firstly, the experimenters make assumptions about some of the features of the "data generating process". For instance, in the study protocol, they stipulate a "true" vaccine efficacy $\widehat{\text{VE}} = 0.6$ (i.e. a reduction by 40% in the infection rate in the vaccine arm, in comparison to the placebo population). This implies that the "true" proportion of COVID-19 cases in the vaccine arm over the total of cases is

$$\hat{\theta} = \frac{1 - \widehat{\text{VE}}}{2 - \widehat{\text{VE}}} = 0.2857.$$

Using these assumption, we can simulate a large number S (say, 100,000) of potential trial data from the alleged generating process $y^{(s)} \sim \text{Bin}(\hat{\theta}, n)$, where the superscript (s) indicates the $s-$th simulated dataset and given a fixed value of the overall number of cases n. Typically, we repeat the simulation for a grid of possible values of n (e.g. $n = [10, 20, 30, 40, 50, 60, 70, \ldots]$).

Once the hypothetical trial data have been generated, the second step consists in analysing them according to the statistical analysis plan defined in the protocol. In this case, the full model specification is required, which involves defining a prior distribution for θ.

For simplicity, the experimenters set a minimally informative Beta prior $\theta \sim \text{Beta}(\alpha_0, \beta_0)$, where the parameters α_0 and β_0 are selected to express the limited amount of information available a priori (recall that θ represents the proportion of COVID-19 cases occurred in the vaccine group). We know that $\text{E}[\theta] = \frac{\alpha_0}{\alpha_0 + \beta_0}$. If we fix $\beta_0 = 1$, then we can solve for α_0 so that the prior mean for θ is equal to some pre-specified value — in particular, the experimenters had chosen a threshold of $\text{VE} = 30\%$ as the minimum level of efficacy they were prepared to entertain. This value can be mapped to the scale of θ as $\frac{1 - 0.3}{2 - 0.3} = 0.4117$ and thus solving

$$\frac{\alpha_0}{\alpha_0 + 1} = 0.4117$$

gives $\alpha_0 = 0.700102$.

Because of conjugacy, for each simulated dataset we can easily update the Beta$(0.700102, 1)$ prior to a Beta(α_1, β_1) posterior distribution for θ, where $\alpha_1 = 0.700102 + y^{(s)}$ and $\beta_1 = 1 + n - y^{(s)}$. Moreover, for each simulation s, we can compute *analytically* any tail-area probability from $p(\theta \mid \text{data})$. Once again, we are really interested in $\Pr(\text{VE} > 0.3 \mid \text{data})$, but using the deterministic relationship linking VE to θ, we can re-express this as

$$\begin{aligned}\Pr(\text{VE} > 0.3 \mid \text{data}) &= \Pr\left(\frac{1-2\theta}{1-\theta} > 0.3 \mid \text{data}\right) \\ &= \Pr(\theta < 0.4117 \mid \text{data}),\end{aligned}$$

which can be computed for each simulation s. This produces a large number of simulations that can be used to determine the "power" for a given sample size, as the proportion of times in which this computed probability exceeds a set threshold (i.e. is "large enough" in the phrasing of the study protocol).

The experimenters compute $n = 164$ as the optimal number of total COVID-19 cases that are necessary to be able to ascertain that $\text{VE} > 30\%$ with a large probability.

Then, it is necessary to determine the overall sample size (i.e. the total number of individuals to be recruited in the study) so that a total of 164 cases is likely to be observed within the required time frame. Once again, it is necessary to make some assumption about the data generating process; specifically the experimenters consider a 1.3% illness rate per year in the placebo group. Because the study aims at accruing 164 cases within 6 months, this essentially amounts to assuming that $\pi_{\text{plac}} \approx 0.013/2$ and thus $\pi_{\text{vac}} \approx (\pi_{\text{plac}} \times 0.4)/2$ (recall that we are assuming a 60% vaccine efficacy, or that $\pi_{\text{vac}} = 0.4 \times \pi_{\text{plac}}$).

We can once again resort to simulations to estimate what sample size in each arm N is necessary so that we can expect $y_{\text{vac}} + y_{\text{plac}} \geq 164$ — this returns an optimal sample size of $N = 17\,600$ per group. Finally, considering an attrition rate of 20% (indicating that such proportion of individuals would not generate an evaluable outcome could be observed), the experimenters inflate the sample size to obtain $N^* = \dfrac{N}{0.8} = 21\,999$ per group (or a total of 43,998 individuals).

Data analysis

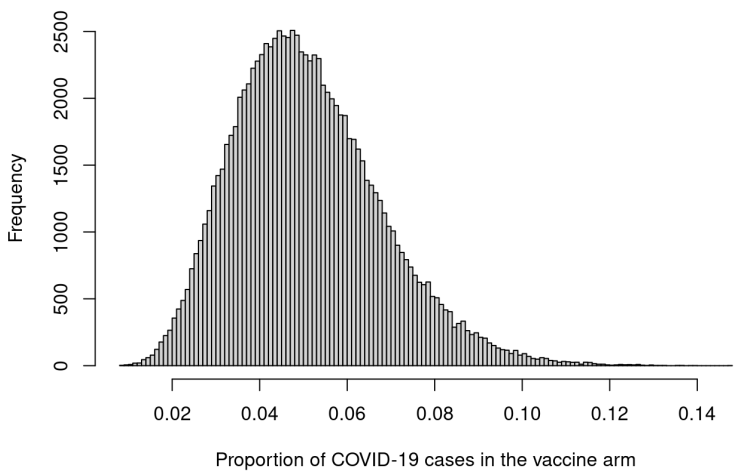
At the end of the actual study, there were $y_{\text{vac}} = 8$ and $y_{\text{plac}} = 162$ confirmed COVID-19 cases in the vaccine and the placebo group, respectively. These imply that $y = 8$ and $n = (8 + 162) = 170$. However, there is a slight extra complication: the setup describe above implies the assumption of 1:1 allocation (i.e. $n_{\text{plac}} = n_{\text{vac}}$) — this is crucial to ensure the simple deterministic relationship between VE and θ. In actual fact, the placebo group had a slightly higher number of individuals, by the time of the data analysis (there were 17,411 and 17,511 individuals with valid data in the vaccine and placebo group, respectively). Thus, we need to rescale the observed data, which can be done by considering $y^*_{\text{vac}} = y_{\text{vac}} \dfrac{17\,461}{17\,411} = 8.02297$ and $y^*_{\text{plac}} = y_{\text{plac}} \dfrac{17\,461}{17\,511} = 161.53743$, where $17\,461 = \dfrac{17\,411 + 17\,511}{2}$ and thus $y^* = 8.02297$ and $n^* = 169.5604$.

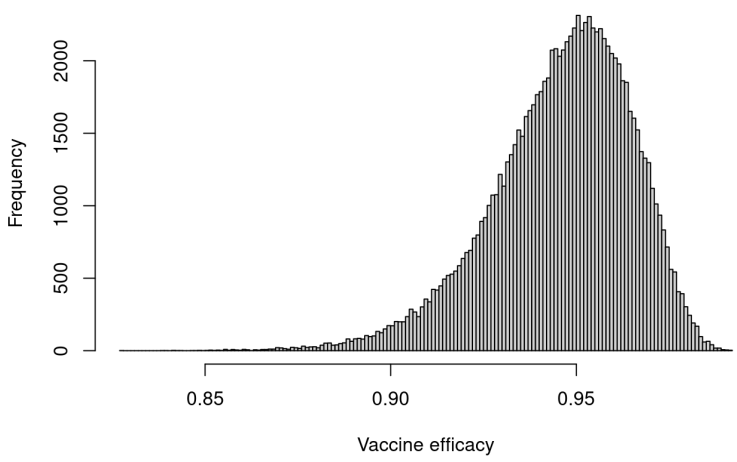
These data can be used to update the prior distribution Beta$(0.700102, 1)$ into a Beta$(8.723072, 162.5374)$ posterior for θ. Finally, we can rescale this posterior to determine the relevant posterior distribution $p(\text{VE} \mid \text{data})$, for instance using Monte Carlo to simulate a large number S of values $\theta^{(s)} \sim \text{Beta}(8.723072, 162.5374)$ and then computing $\text{VE}^{(s)} = \dfrac{1-2\theta^{(s)}}{1-\theta^{(s)}}$ and then use these to characterise the uncertainty around the vaccine efficacy.

In this case, the posterior distributions $p(\theta \mid \text{data})$ and $p(\text{VE} \mid \text{data})$ can be visualised below.

```
alpha.0=0.700102
beta.0=1
y=c(8,162)
n=c(17411,17511)
# So needs to reproportion as if the treatment arms had the same sample size (to be in line with mc
y=y*mean(n)/n
# Now can update the Beta prior with the observed data
alpha.1=alpha.0+y[1]
beta.1=beta.0+y[2]

theta=rbeta(100000,alpha.1,beta.1)
ve=(1-2*theta)/(1-theta)
```





We can also use the resulting samples from the posterior distribution to estimate the 95% interval of $[0.9034; 0.9761]$ for the vaccine efficacy, indicating that we expect the vaccine to perform extremely well.

PREVIOUS
Practical 2. Markov Chain Monte Carlo - SOLUTIONS

NEXT
How do MCMC and Gibbs sampling really work?