# Optimum Simulations

wP vs aP

*Prepared by: James Totterdell*

*2019-02-11*

## Contents

## 1 Background and Rationale

The aim is to o assess the allergy-preventive benefit and the safety of using wP as the first infant pertussis vaccine dose, compared with using aP for all doses.

## 2 Primary Outcome

The primary outcome is challenge-proven IgE-mediated food allergy by age 18-months.

The end-point is challenge-proven IgE-mediated food allergy at 18-months.

# 3 Sample size and accrual

The end-point is food allergy at 18-months. For simplicity, assume babies are enrolled and randomised at 0 months of age. So, no follow-up data is available until 18-months after the first infant is enrolled.

Suppose accrual is 20 infants per week, then, by the time we have follow-up on the first individual we will have enrolled 1,560 infants (78 weeks × 20 per week). Assuming the first analysis was at $n = 500$, we would have about 2,000 individuals enrolled, so 1,500 with missing information at the time of the first interim. Full follow-up would occur at about week 228 (Figure 1).

Suppose accrual is 10 infants per week, then, by the time we have follow-up on the first individual we will have enrolled 780 infants. Assuming the first analysis was at $n = 500$, we would have about 1,300 enrolled, so 800 with missing information at the first interim. Full follow-up would occur at about week 378.

The minimum acrrual rate needed to enroll 3,000 infants over 5 years is about 11.5 per week.

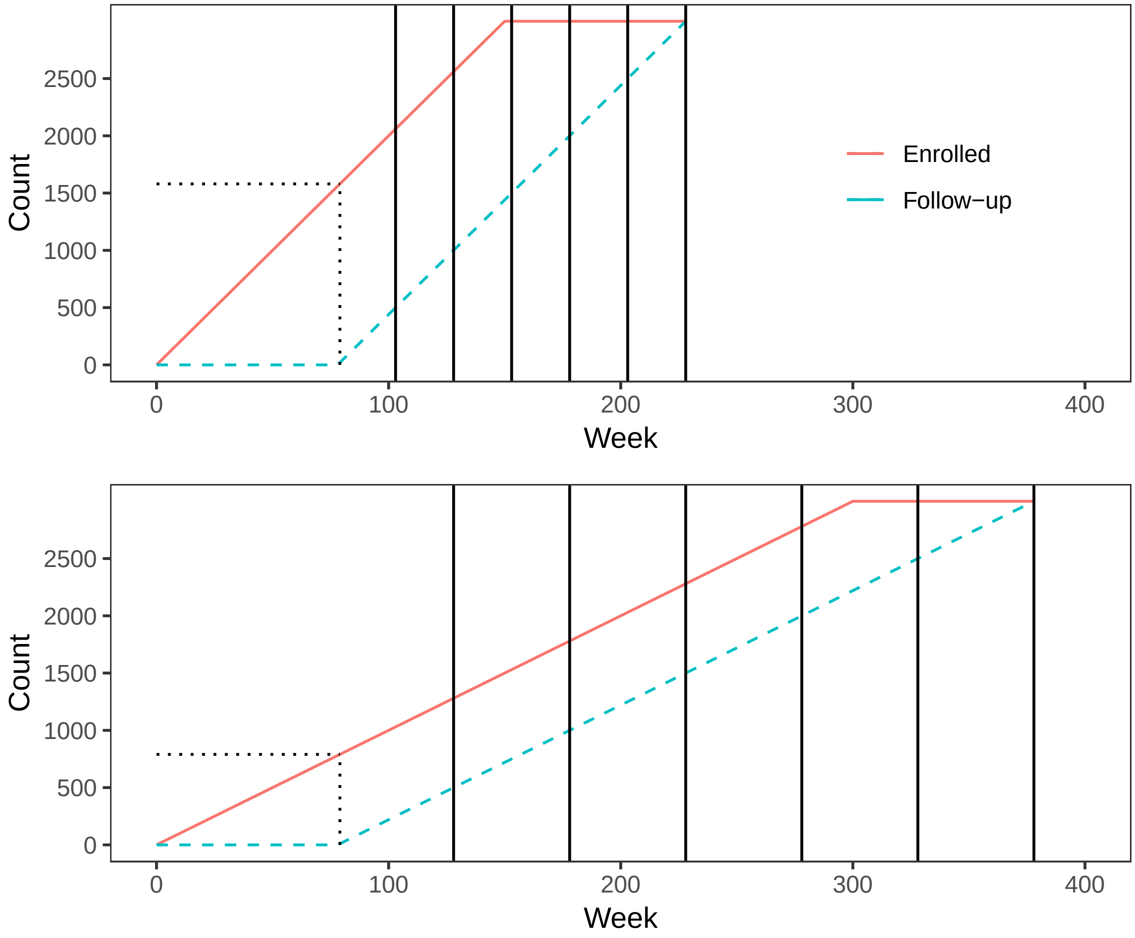

Figure 1: Assumed accrual rate and associated delay of information.

# 4 Statistical Analysis

## 4.1 Model

Let $\theta_a$ be the proportion of infants with food allergy who received the acellular pertussis vaccine, and $\theta_w$ the proportion of infants wit hfood allergy who received the whole-cell pertussis vaccine. We are interested in estimating $\delta = \theta_w - \theta_a$ and our hypothesis corresponds to

$$
\begin{aligned}
H_0 &: \delta \geq 0 \\
H_1 &: \delta < 0
\end{aligned}
$$

That is, that $\theta_w$ is no improvement over $\theta_a$ versus $\theta_w$ is lower than $\theta_a$.

We could approach this in two ways:

- We might model $\theta_a$ and $\theta_w$ directly using independent Beta-Binomial models and integrate over the random variable $\delta$ to obtain posterior probabilities. The advantage is simplicity.
- We might model $\theta_a = 1/(1 + \exp(\beta_0)$, and $\theta_b = 1/(1 + \exp(-\beta_0 - \beta_1))$, that is as a logistic regression model, where $\delta = \beta_2$ is our parameter of interest (the difference in log-odds). The advantage we can incorporate more complexity (subgroups, partial pooling etc).

Given the lengthy delay in information, we will likely utilise posterior predictive probabilities at any interim analyses to impute the as yet unobserved data.

## 4.2 Independent Beta-Binomial Models

Suppose that at each analysis $k = 1, ..., K$ we have data on $n_k^i$ individuals with $y_k^i$ responses for $i \in \{a, w\}$. We also assume that we have $m_k^i \geq n_k^i$ total enrolled but not all with data. The number without data is $\tilde{n}_k^i = m_k^i - n_k^i$. At an interim analysis we wish to impute the data for individuals enrolled but without follow-up. We denote these missing number of responses as $\tilde{y}_k^i$.

In addition to enrolled individuals with missing data, there are the yet to be enrolled individuals making up the maximum sample size. At stage $K$ we have $n_K^i$ individuals with $y_K^i$ responses, and so for this end point we have $\tilde{n}_k^i = n_K^i - n_k^i$ data points missing. In either case, the posterior predictive will have the same parameters but with a different sample size parameter. Therefore in what follows we do not distinguish between the two, however, it is standard to use $\tilde{n}_k^i = m_k^i - n_k^i$ in deciding success and $\tilde{n}_k^i = n_K^i - n_k^i$ in deciding futility.

We specify the following model for $i \in \{a, w\}$ and $k \in \{1, ..., K\}$,

$$
\begin{aligned}
\pi_0^i(\theta^i) &= \text{Beta}(\theta^i | a^i, b^i) \\
f_k^i(y_k^i | \theta^i) &= \text{Binomial}(n_k^i, y_k^i) \\
\pi_k^a(\theta^i | y_k^i) &= \text{Beta}(\theta^i | a^i + y_k^i, b^i + n_k^i - y_k^i) \\
P_k &= \mathbb{P}_{\Theta^a, \Theta^w | Y_k^a, Y_k^w}(\theta^w < \theta^a) \\
&= \int_0^1 \pi_k^a(\theta^a | y_k^a) \left[ \int_0^{\theta^a} \pi_k^w(\theta^w | y_k^w) d\theta^w \right] d\theta^a \\
\tilde{f}_k^i(\tilde{y}_k^i | y_k^i) &= \text{Beta-Binomial}(\tilde{y}_k^i | \tilde{n}_k^i, a^i + y_k^i, b^i + n_k^i - y_k^i) \\
\text{PPoS}_k &= \mathbb{E}_{\tilde{Y}_k^a, \tilde{Y}_k^w | Y_k^a, Y_k^w} \left[ \mathbb{I} \left\{ \mathbb{P}_{\Theta^a, \Theta^w | Y_k^a + \tilde{Y}_k^a, Y_k^w + \tilde{Y}_k^w}(\theta^w < \theta^a) > k \right\} \right] \\
&= \sum_{i=0}^{\tilde{n}_k^a} \sum_{j=0}^{n_k^w} \mathbb{I} \left\{ \int_0^1 \tilde{\pi}_k^a(\theta^a | y_k^a + i) \left[ \int_0^{\theta^a} \tilde{\pi}_k^w(\theta^w | y_k^w + j) d\theta^w \right] d\theta^a > q \right\} \tilde{f}_k^w(j | y_k^w) \tilde{f}_k^a(i | y_k^a)
\end{aligned}
$$

The quantity $P_k$ cannot be calculated analytically but can be evaluated numerically or estimated using Monte Carlo methods. Although $\text{PPoS}_k$ can be computed analytically (assuming we have calculated the relevant $\tilde{P}_k$) it may still be more efficient to estimate using Monte Carlo methods for large sample sizes.

Other options for speeding things up is to pre-determine all possible $\tilde{P}_k$ values for the possible values of $i, j$, or to determine which $i, j$ have probability greater than some minimum threshold (e.g. $10^{-8}$) and only determine $\tilde{P}_k$ for that subset of possibilities.

## 4.3 Logistic Regression

Advantage is that we can incorporate additional covariates. However it complicates the simulations. Posterior probabilities are no longer analytically tractable, but instead must be approximated by Monte Carlo measures, or by approximating densities.

We now specify

$$\begin{aligned}
\theta_k^a &= \text{logit}^{-1}(\beta_0) \\
\theta_k^w &= \text{logit}^{-1}(\beta_0 + \beta_1) \\
f_k(y_k^a|\theta_k^a) &= \text{Binomial}(y_k^a|n_k^a, \theta_k^a) \\
f_k(y_k^w|\theta_k^w) &= \text{Binomial}(y_k^w|n_k^w, \theta_k^w) \\
\pi_k(\theta_k^a|y_k^a) &\approx N^{-1} \sum_{i=1}^{N} \delta_{\theta_k^{a,i}}(d\theta_k^a), \quad \{\theta_k^{a,i}\}_{i=1}^N \sim \pi_k(\theta_k^a|y_k^a) \\
\pi_k(\theta_k^a|y_k^a) &\approx \sum_{i=1}^{N} \delta_{\theta_k^{a,i}}(d\theta_k^a)
\end{aligned}$$

# 5 Simulations

Parameters we need to configure are:

- $\underline{c}_K, \overline{c}_K$ - the bounds used at the final analysis for decisions
- $q_k$ - the bounds used for determining posterior probability of success in $\text{PPos}_k$ (should this just be $\overline{c}_K$?)