

# Informative Prior Elicitation Using Historical Data, Expert Opinion, and Other Sources

Ethan M. Alt<sup>1</sup>

Department of Biostatistics, UNC Chapel Hill

Joseph G. Ibrahim

Department of Biostatistics, UNC Chapel Hill

---

<sup>1</sup>Dr. Matthew Psioda assisted with the development of these notes

## Part 1: Introduction to the Bayesian Paradigm

## Introduction

# Advantages of Bayesian Methods in Clinical Trials

- The Bayesian expresses uncertainty about an unknown parameter  $\theta$  using an entire distribution as opposed to a point estimate and a standard error.
- Interpretations of Bayesian quantities are more intuitive.
  - ▶ One cannot make a statement such as “the mean  $\theta$  has a 95% chance of being in the interval  $\bar{x} \pm 1.96s/\sqrt{n}$ ”, although many are tempted.
  - ▶ This is precisely the interpretation of a Bayesian 95% credible interval!
  - ▶ For a frequentist, “95%” is only a label that indicates how the interval estimation procedure performs over the long haul.
- Bayesian inference follows the likelihood principle: any two sampling designs leading to the same likelihood function lead to identical Bayesian inferences.

## Advantages of Bayesian Methods in Clinical Trials

- Easy to make predictions using posterior predictive distributions.
- Bayesian methods dovetail naturally into decision analysis and cost-effectiveness analysis.
- Bayesian methods allow for flexible and complex designs that exhibit good operating characteristics.
  - ▶ Examples of this include adaptive designs, flexible interim monitoring, and designs for complex models such as multivariate survival models, meta-analysis models, and models for longitudinal data.
- The Bayesian paradigm provides a natural mechanism for the incorporation of prior information and historical data.

## The Bayes paradigm

# The Bayes paradigm

- We discuss the Bayesian paradigm for **parametric** inference.
- Suppose we conduct (or design) a study, in which the parameter  $\theta$  is of interest. For example,
  - ▶  $\theta$  = difference in means
  - ▶  $\theta$  = hazard ratio
  - ▶  $\theta$  = vector of regression coefficients
  - ▶  $\theta$  = probability that a treatment is effective
- Letting  $y_i : i = 1, \dots, n$  denote the data, we denote the **sampling distribution for the data** by  $p(y_i|\theta)$ .
- The **likelihood**, denoted  $\mathcal{L}(\theta|\mathbf{y})$ , is any function that satisfies

$$\mathcal{L}(\theta|\mathbf{y}) \propto \prod_{i=1}^n p(y_i|\theta).$$

# The Bayes Paradigm

- Suppose  $y_i \mid \theta \sim \text{Bernoulli}(\theta)$ . Then,

$$\mathcal{L}(\theta | \mathbf{y}) \propto \theta^{\sum y_i} (1 - \theta)^{n - \sum y_i}.$$

- In the Bayesian mindset, we express our uncertainty about unknown quantities by specifying distributions for them.
- Thus, we express our uncertainty about  $\theta$  by specifying a **prior distribution** for it, denoted by  $\pi(\theta)$ .
- Using Bayes' theorem, we combine the likelihood and prior distribution to construct the distribution of  $\theta | \mathbf{y}$ , called the **posterior distribution** and denoted by  $p(\theta | \mathbf{y})$ .

## Bayes theorem

- By Bayes Theorem,

$$p(\theta|\mathbf{y}) = \frac{\mathcal{L}(\theta|\mathbf{y}) \pi(\theta)}{\int_{\Theta} \mathcal{L}(\theta|\mathbf{y}) \pi(\theta) d\theta} ,$$

where  $\Theta$  denotes the parameter space of  $\theta$ .

- The quantity

$$m(\mathbf{y}) = \int_{\Theta} \mathcal{L}(\theta|\mathbf{y}) \pi(\theta) d\theta$$

is the **normalizing constant** of  $p(\theta|\mathbf{y})$ .

- $m(\mathbf{y})$  is also called the **marginal likelihood**, **evidence**, or **model evidence**.
- Note that  $m(\mathbf{y})$  depends only on the data  $\mathbf{y}$  and the prior  $\pi(\cdot)$ , but not the parameter  $\theta$ .

## Bayes theorem

- For most inference problems,  $m(\mathbf{y})$  does not have a closed form but its computation is not critical for Bayesian inference due to **Markov Chain Monte Carlo (MCMC)** methods for model fitting [1].
- Bayesian inference about  $\theta$  is primarily based on the posterior distribution of  $\theta$ , denoted by  $p(\theta|\mathbf{y})$ .
- One can compute various posterior summaries for  $\theta$  (or any function of  $\theta$ ), such as the mean, median, variance, and quantiles.
  - ▶ The **posterior mean** of  $\theta$  is given by  $E(\theta | \mathbf{y}) = \int_{\Theta} \theta \pi(\theta | \mathbf{y}) d\theta$ .
  - ▶ The **maximum a posteriori (MAP)** estimate (i.e., posterior mode) is the value of  $\theta$  that maximizes  $\pi(\theta | \mathbf{y})$ .

## Example: Bernoulli proportion model I

- Suppose we possess data  $\{y_i, i = 1, \dots, n\}$  where  $y_i \in \{0, 1\}$ .
- The likelihood may be expressed as

$$\mathcal{L}(\theta | \mathbf{y}) \propto \theta^{n\bar{y}} (1 - \theta)^{n(1-\bar{y})}$$

- Consider a beta prior of the form

$$\pi(\theta) = \frac{1}{B(\alpha_0, \beta_0)} \theta^{\alpha_0-1} (1 - \theta)^{\beta_0-1}$$

- The posterior density is a beta distribution

$$p(\theta | \mathbf{y}) \propto \mathcal{L}(\theta | \mathbf{y}) \pi(\theta) \propto \theta^{n\bar{y} + \alpha_0 - 1} (1 - \theta)^{n(1-\bar{y}) + \beta_0 - 1}$$

- When the prior and posterior are of the same family of distributions, we say that the prior is a **conjugate** prior.

## Example: Bernoulli proportion model II

- The posterior mean is given by

$$E(\theta|\mathbf{y}) = \frac{n\bar{y} + \alpha_0}{n + \alpha_0 + \beta_0} = \frac{\bar{y}}{1 + \frac{\alpha_0 + \beta_0}{n}} + \frac{\frac{\alpha_0}{n}}{1 + \frac{\alpha_0 + \beta_0}{n}}.$$

- Note as  $n \rightarrow \infty$  and/or as  $\alpha_0, \beta_0 \rightarrow 0$ ,  $E(\theta|\mathbf{y}) \rightarrow \bar{y}$ , which is the MLE.
- It can be shown that the MAP estimate is given by

$$\hat{\theta}_{\text{MAP}} = \frac{n\bar{y} + \alpha_0 - 1}{n + \alpha_0 + \beta_0 - 2}$$

- When  $\alpha_0 = \beta_0 = 1$  (i.e., a uniform prior for  $\theta$ ),  $\hat{\theta}_{\text{MAP}} = \bar{y}$

## Predictive distributions

## Making predictions

- Given the observed data  $\mathbf{y}$ , we carry out predictions by computing the Bayesian predictive distribution, or **posterior predictive distribution**.
- The posterior predictive distribution of future response values  $\mathbf{y}^*$  is defined as

$$p(\mathbf{y}^*|\mathbf{y}) = \int_{\Theta} p(\mathbf{y}^*|\theta)p(\theta|\mathbf{y}) d\theta,$$

where  $p(\theta|\mathbf{y})$  is the posterior distribution for  $\theta$  based on  $\mathbf{y}$  alone.

- Predictive distributions are fundamental to Bayesian sample size determination and trial monitoring.

## Example: Bernoulli model

- The posterior predictive distribution for a single observation is given by

$$\begin{aligned} p(y^* | \mathbf{y}) &= \int \theta^{y^*} (1 - \theta)^{y^*} \frac{\theta^{n\bar{y} + \alpha_0 - 1} (1 - \theta)^{n(1 - \bar{y}) + \beta_0 - 1}}{B(n\bar{y} + \alpha_0, n(1 - \bar{y}) + \beta_0)} \\ &= \frac{B(y^* + n\bar{y} + \alpha_0, 1 - y^* + n(1 - \bar{y}) + \beta_0)}{B(\bar{y} + \alpha_0, n(1 - \bar{y}) + \beta_0)}, \end{aligned}$$

which can be shown to be a Bernoulli distribution with probability of success  $p^* = \frac{n\bar{y} + \alpha_0}{n + \alpha_0 + \beta_0}$

- Note that we can alternatively sample from this distribution via the following hierarchical scheme:
  - Sample  $\theta | \mathbf{y} \sim \text{Beta}(n\bar{y} + \alpha_0, n(1 - \bar{y}) + \beta_0)$
  - Sample  $y^* | \theta, \mathbf{y} \sim \text{Ber}(\theta)$

Bayesian interval estimation

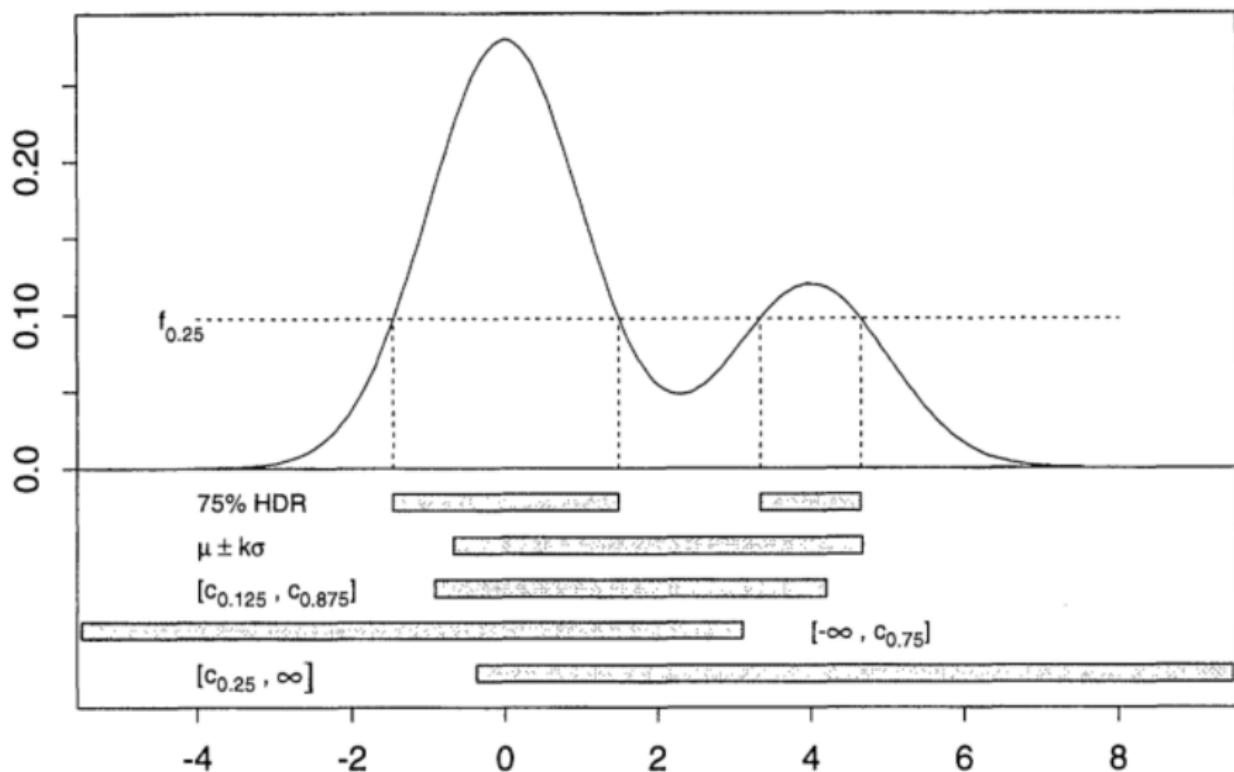
## Bayesian interval estimation

- A  $(1 - \alpha) \times 100\%$  **credible interval** (CI) is any interval  $(a, b)$  that satisfies

$$\Pr(\theta \in (a, b) | \mathbf{y}) = \int_a^b p(\theta | \mathbf{y}) d\theta = 0.95$$

- We interpret a CI as follows: there is a 95% chance (or probability) that the true parameter  $\theta$  is between  $(a, b)$ .
- A special case of a CI is called a **highest posterior density**(HPD) intervals, which is essentially defined to be a CI with minimal width.
- When the posterior density is symmetric and unimodal, HPD intervals are easy to calculate.
- An easy way to obtain a CI is to use the  $\alpha/2$  and  $1 - \alpha/2$  percentiles of the posterior density.

# Bayesian interval estimation



## Bayesian hypothesis testing

# Bayesian hypothesis testing I

- Canonical Bayesian hypothesis testing involves imposing priors on the null and alternative hypotheses and deriving the **Bayes factor** (BF), namely,

$$BF = \frac{\text{Posterior odds}}{\text{Prior odds}} = \frac{p(H_1|\mathbf{y})/p(H_0|\mathbf{y})}{\pi(H_1)/\pi(H_0)},$$

where  $p(H_0|\mathbf{y}) = \frac{m(\mathbf{y}|H_0)}{m(\mathbf{y}|H_0) + m(\mathbf{y}|H_1)}$  and

$$m(\mathbf{y}|H_a) = \int_{\theta \in H_a} \mathcal{L}(\theta|\mathbf{y})\pi(\theta|H_a)d\theta$$

is the **marginal likelihood** (or evidence) for hypothesis  $H_a$ .

- The **posterior model probability** of  $H_0$  is given by

$$p(H_0|\mathbf{y}) = \frac{m(\mathbf{y}|H_0)}{m(\mathbf{y}|H_0) + m(\mathbf{y}|H_1)}$$

## Bayesian hypothesis testing II

- In this course, we focus on posterior probabilities of half-spaces to conduct hypothesis testing.
- In particular, if  $\theta$  is a log-hazard ratio, we may compute

$$\Pr(\theta < 0 | \mathbf{y}) = \int_{-\infty}^0 p(\theta | \mathbf{y})$$

- **Theorem:** If the data,  $\mathbf{y}$  were generated from  $H_0 : \theta = 0$ , then

$$\Pr(\theta < 0 | \mathbf{y}) \xrightarrow{d} U(0, 1) \text{ as } n \rightarrow \infty.$$

This is referred to as the **asymptotic uniformity** of the posterior probability over the half-space under the null hypothesis.

- Thus, it is reasonable to reject  $H_0$  if  $\Pr(\theta < 0 | \mathbf{y}) > 1 - \alpha$ , which ensures a type I error rate of  $\alpha$  in large samples.

## Bayesian multiple hypothesis testing

- In clinical trials, it is common to have multiple endpoints.
- Suppose  $y_1 = \text{reduction in systolic blood pressure}$  and  $y_2 = \text{reduction in diastolic blood pressure}$  with mean and covariance parameters  $\mu = (\mu_1, \mu_2)'$  and  $\Sigma$ .
- Alt et al. (2023) [2] showed that, if the null hypothesis  $H_0 : \mu_1 = \mu_2 = 0$  is true, then

$$\Pr(\mu_1 > 0 \text{ or } \mu_2 > 0 | \mathbf{y}) \rightarrow 1 - \Phi_2(\mathbf{Z} | \mathbf{0}, \boldsymbol{\Gamma}^*), \quad \mathbf{Z} \sim N_2(\mathbf{0}, \boldsymbol{\Gamma}^*), \quad (1)$$

where

- ①  $\Phi_J(\cdot | \mu, \mathbf{C})$  is the  $J$ -dimensional multivariate normal CDF with mean  $\mu$  and covariance matrix  $\mathbf{C}$ .
- ②  $\boldsymbol{\Gamma}^*$  is the **limiting** posterior correlation matrix of  $\mu_1, \mu_2$  under  $H_0$ .
- Note that if  $J = 1$ , we get back the asymptotic uniformity of the posterior probability.

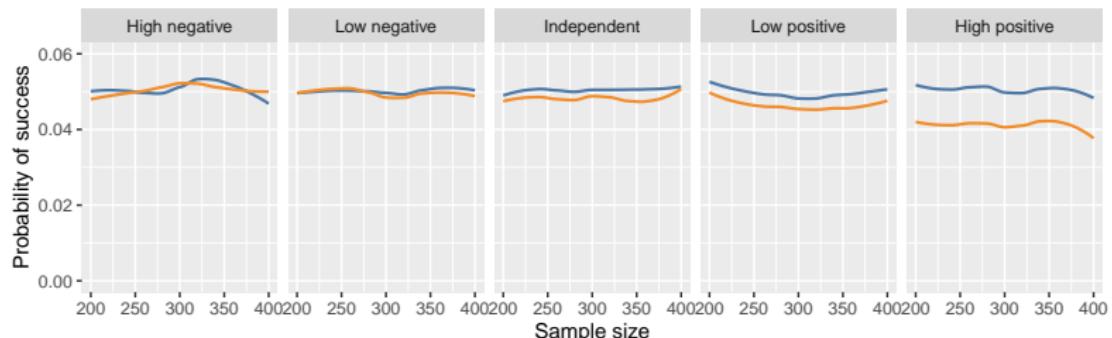
## Bayesian multiple hypothesis testing

- We may sample from (1) via
  - ➊ Sample  $\mathbf{Z} \sim N_J(\mathbf{0}, \boldsymbol{\Gamma})$
  - ➋ Compute  $1 - \Phi_J(\mathbf{Z}|\mathbf{0}, \boldsymbol{\Gamma})$
- Obtaining many samples, we may compute the  $1 - \alpha/2$  quantile of the distribution, say  $P_{1-\alpha/2}$
- We reject the global null  $H_0$  if

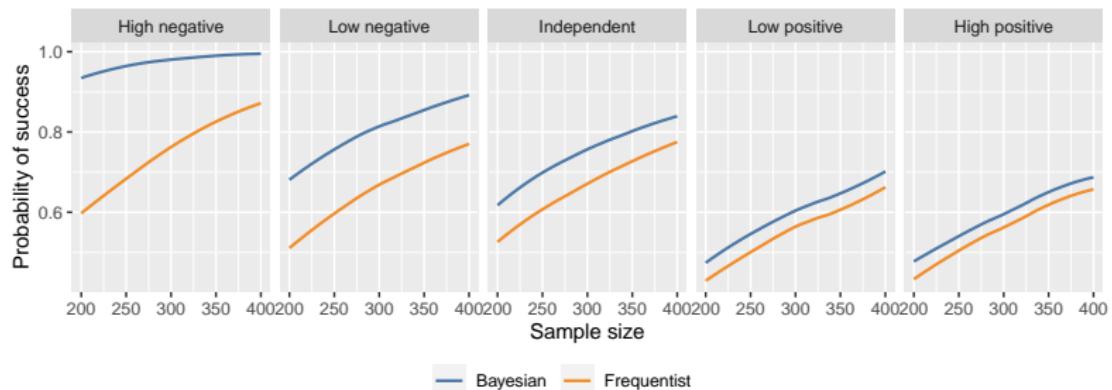
$$\Pr(\mu_1 > 0 \text{ or } \mu_2 > 0 | \mathbf{y}) > P_{1-\alpha/2}$$

# Bayesian multiple hypothesis testing

Type 1 error



BCEP

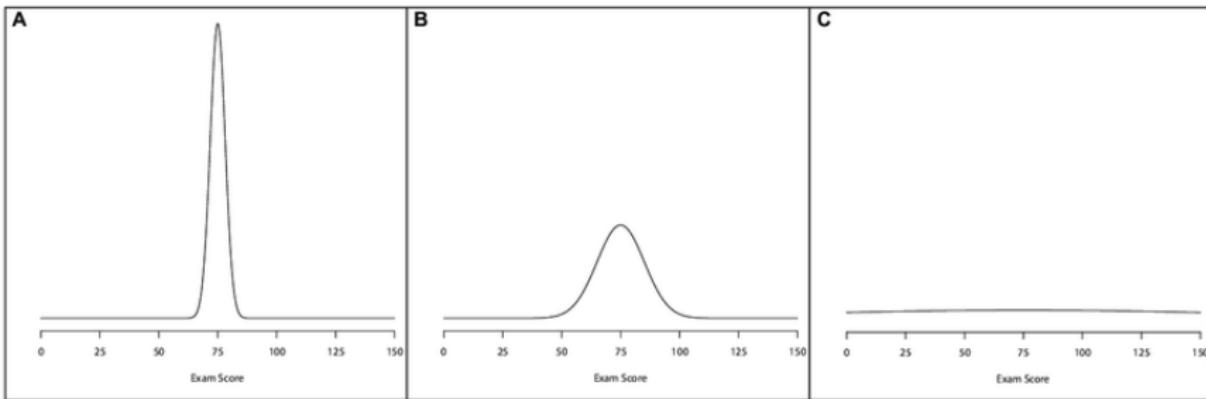


Prior elicitation

## Prior elicitation

- Priors are generally dichotomized into two groups: **informative** and **noninformative** (also called vague or flat).
- An **informative** prior is one that results in a posterior density that is not dominated by the likelihood.
- A **noninformative** prior has a flat shape, and the prior has essentially no impact on the resulting posterior density.
- **Remark:** third group of priors are called **weakly informative** if they are not flat but do not have a big impact on the resulting posterior density.
- Typically, the **prior variance** governs the level of informativeness of the prior.

# Prior elicitation



Examples of (A) = informative prior; (B) = weakly informative prior; (C) = noninformative prior. Source: <http://dx.doi.org/10.3389/fpsyg.2020.608045>

# Priors for exponential families I

- Likelihoods for exponential family models take the form

$$\mathcal{L}(\theta|\mathbf{y}) \propto \exp \left\{ \sum_{i=1}^n (\theta y_i - b(\theta)) \right\} = \exp \{ n(\theta \bar{y} - b(\theta)) \}.$$

- Diagonis and Yvilsaker [3] developed the prior

$$\pi_{DY}(\theta|n_0, \bar{y}_0) \propto \exp \{ n_0[\theta \bar{y}_0 - b(\theta)] \}.$$

- This results in the posterior

$$p(\theta|\mathbf{y}) = \pi_{DY} \left( \theta \middle| n + n_0, \frac{n\bar{y} + n_0\bar{y}_0}{n + n_0} \right).$$

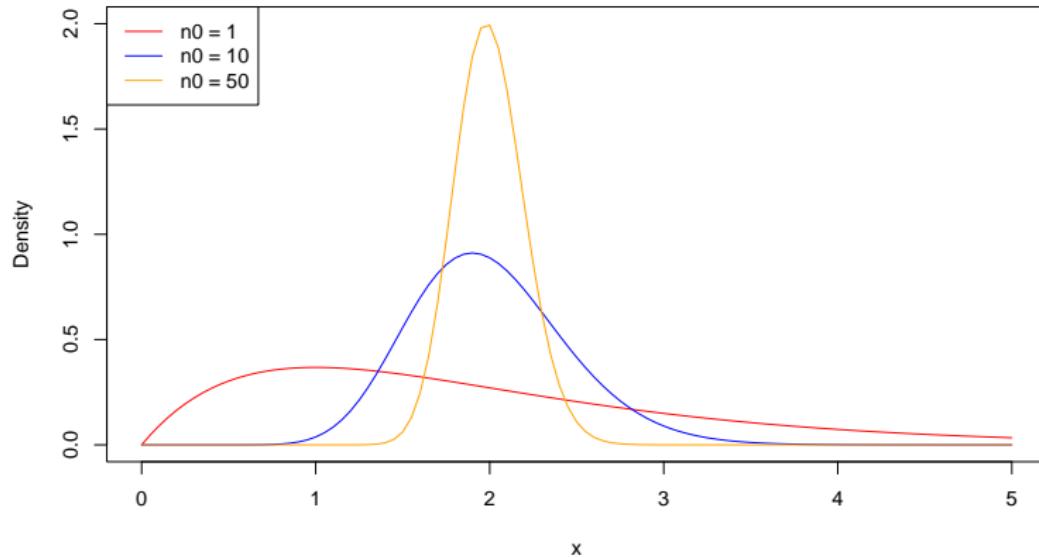
- The posterior **precision** parameter is  $n + n_0$  and the posterior **shape** parameter is a **convex combination** of the mean of the data  $\bar{y}$  and  $\bar{y}_0$ .
- The hyperparameter  $n_0$  is referred to as a **prior sample size** and the hyperparameter  $\bar{y}_0$  is referred to as a **prior prediction** for  $E(y)$ .

## Priors for exponential families II

- The prior precision parameter  $n_0$  determines the level of informativeness of the prior. Typically, we restrict  $n_0 \leq n$
- Note if  $n_0 = 0$ , the prior is a uniform improper prior for  $\theta$ . If  $n_0 = n$ , the prior is given as much weight as the data.
- If summary statistics from a previous study exist, one may plug in  $n_0 = \text{sample size from study}$  and  $\bar{y}_0 = \text{sample mean from study}$ .
- When converted to the scale of the mean, i.e.,  $\mu = b(\theta)$ , the priors are recognizable:

Model	$\theta(\mu)$	Prior ( $\mu$ scale)
$N(\mu, 1)$	$\mu$	$N(\bar{y}_0, n_0^{-1})$
Bernoulli( $\mu$ )	$\log\left(\frac{\mu}{1-\mu}\right)$	Beta( $n_0\bar{y}_0, n_0(1-\bar{y}_0)$ )
Poisson( $\mu$ )	$\log(\mu)$	Gamma( $n_0\bar{y}_0, n_0$ )

# Priors for exponential families III



DY prior for Poisson model when  $\bar{y}_0 = 2$  ( $\mu$ -scale)

## Priors for exponential families IV

- Consider a Bernoulli model with  $\bar{y}_0 = 0.3$  and  $n_0 = 78.8$  and consider  $\gamma = \text{expit}(\theta)$ .
- The DY prior can be expressed as

$$\pi_{\text{DY}}(\theta | \lambda = 78.8, m = 30) \propto \exp \{78.8 [0.30\theta - \log(1 + \exp(\theta))]\}$$

$\downarrow$

$$\gamma = \text{expit}(\theta)$$

$\downarrow$

$$\pi_{\text{DY}}(\gamma | \lambda = 78.8, m = 30) = \text{Beta}(m\lambda = 23.64, (1 - m)\lambda = 55.16)$$

## Noninformative priors

- Loosely speaking, a noninformative is one that does not impact the posterior.
- Noninformative priors may be **proper** or **improper**.
- We may represent any prior of the form

$$\pi(\theta) = \frac{\pi^*(\theta)}{C},$$

where  $C = \int \pi^*(\theta) d\theta$  is a **normalizing constant**

- A prior is **proper** if  $C < \infty$ . A prior is **improper** if  $C = \infty$ .
- A **proper** prior yields a **proper** posterior (with probability 1).
- An **improper** prior **may or may not** yield a **proper** posterior.

## Noninformative priors: improper priors

- Consider an i.i.d. Poisson model (e.g., tumor counts).
- The likelihood is given as

$$L(\lambda|\mathbf{y}) = \prod_{i=1}^n \frac{e^{-\lambda} \lambda^{y_i}}{y_i!} \propto \lambda^{n\bar{y}} e^{-n\lambda}.$$

- Using a conjugate  $\text{Gamma}(\alpha_0, \beta_0)$  prior, we have

$$\pi(\lambda|\alpha_0, \beta_0) \propto \lambda^{\alpha_0-1} e^{-\beta_0 \lambda} \mathbf{1}\{\lambda > 0\}$$

the posterior is given by

$$p(\lambda|\mathbf{y}) \propto \lambda^{n\bar{y} + \alpha_0 - 1} e^{-(n + \beta_0)\lambda}.$$

- Note that if  $\alpha_0 = 1, \beta_0 = 0$ ,
  - ▶ The posterior is still proper!  $\lambda|\mathbf{y} \sim \text{Gamma}(n\bar{y}, n)$ .
  - ▶ The prior is an **improper uniform prior** over  $\lambda \in (0, \infty)$ .

## Noninformative priors: improper priors

- We see that improper priors **can** lead to proper posteriors.
- Suppose we have a single observation  $y \sim \text{Ber}(\theta)$ .
- We elicit  $\pi(\theta) \propto \theta^{-1}(1 - \theta)^{-1} = \frac{1}{\theta} + \frac{1}{1-\theta}$ .
- Note that  $\int_0^1 \pi(\theta) d\theta = [\log(\theta)]_{\theta=0}^1 - [\log(1 - \theta)]_{\theta=0}^1 = \infty$ .
- The posterior is

$$p(\theta|y) \propto \theta^{y-1}(1 - \theta)^{-y} = \begin{cases} \frac{1}{1-\theta} & \text{if } y = 1 \\ \frac{1}{\theta} & \text{if } y = 0 \end{cases},$$

which is easily shown to be improper for any  $y \in \{0, 1\}$ .

- When in doubt about posterior propriety, use a **proper prior with large variance** (e.g., a DY prior with  $n_0 \approx 0$ ).

## Noninformative priors: Jeffreys' prior

- Jeffreys' prior is a popular noninformative prior given as

$$\pi_J(\boldsymbol{\theta}) \propto |\mathcal{I}(\boldsymbol{\theta})|^{1/2},$$

where  $\mathcal{I}(\boldsymbol{\theta}) = E_{\mathbf{y}|\boldsymbol{\theta}} \left[ -\frac{\partial^2 \log \mathcal{L}(\boldsymbol{\theta}|\mathbf{y})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \right]$  is the Fisher information matrix.

- For example, suppose  $y_i \sim \text{Bernoulli}(p)$ . Then

$$\mathcal{I}(p) = E_{\mathbf{y}|p} \left[ \frac{n\bar{y}}{p^2} + \frac{n(1-\bar{y})}{(1-p)^2} \right] = \frac{n}{p(1-p)} \propto p^{-1}(1-p)^{-1}$$

- Thus, Jeffreys' prior is given by

$$\pi_J(p) \propto p^{-1/2}(1-p)^{-1/2} \propto \text{Beta}(p|0.5, 0.5),$$

which is proper!

## Noninformative priors: Jeffreys' prior

- Jeffreys' prior is **invariant** under a change of coordinates.
- Let  $y \sim \text{Ber} \left( \frac{\theta}{1+\theta} \right)$  (odds parameterization).

$$f(y|\theta) = \left( \frac{\theta}{1+\theta} \right)^y \left( \frac{1}{1+\theta} \right)^{1-y}$$

- Jeffreys' prior for  $\theta$  is

$$\pi_J(\theta) \propto [\theta^{-1}(1+\theta)^{-2}]^{1/2} = \theta^{-1/2}(1+\theta)^{-1}$$

- Now, note that  $\frac{\partial \theta}{\partial p} = (1-p)^{-2}$ , and Jeffreys' prior converting  $\theta$  to  $p$  is

$$\pi_J(\theta \rightarrow p) \propto p^{-1/2}(1-p)^{-1/2},$$

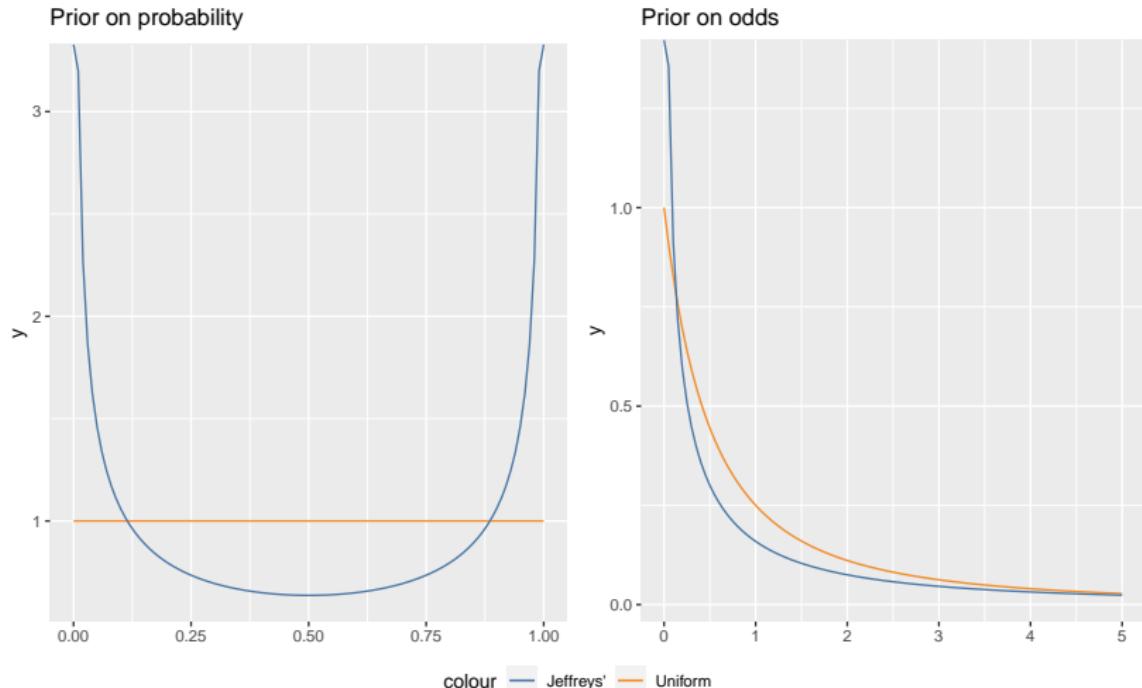
which is Jeffreys' prior obtained from starting with the  $p$  parameterization.

## Noninformative priors: Jeffreys' prior

- It also turns out that Jeffreys' prior is noninformative in the sense of maximizing the distance between prior and posterior.
- The facts that Jeffreys' prior is (1) noninformative and (2) invariant to transformations make it a nice default option.
- E.g., under any reparameterization, Jeffreys' prior is noninformative.
- Consider a  $\text{Ber}(p)$  model. A reasonable noninformative prior is  $p \sim U(0, 1) = \text{Beta}(1, 1)$ . This prior in terms of the odds,  $\theta$ , is given by

$$\pi_{U(0,1)}(\theta) = (1 + \theta)^{-2}, \quad \theta > 0$$

# Noninformative priors: Jeffreys' prior



Uniform and Jeffreys' priors for Bernoulli model.

## Jeffreys' prior for a normal linear model

- For a normal linear model, the Fisher information matrix is given by

$$\mathcal{I}(\beta, \sigma^2) = \begin{pmatrix} \frac{\mathbf{X}'\mathbf{X}}{\sigma^2} & 0 \\ 0 & \frac{n}{\sigma^4} \end{pmatrix}.$$

- If  $\sigma^2$  is known, then

$$\pi_J(\beta | \sigma^2) \propto |\sigma^{-2} \mathbf{X}'\mathbf{X}|^{1/2} \propto 1.$$

- If  $\sigma^2$  is unknown, then

$$\pi_J(\beta, \sigma^2) \propto |\sigma^{-2} \mathbf{X}'\mathbf{X}|^{1/2} \left( \frac{n}{\sigma^4} \right)^{1/2} \propto (\sigma^2)^{-\frac{p+2}{2}} = \sigma^{-p/2+1}.$$

- In either case, the prior is improper, but the resulting posterior is proper.

## Jeffreys' priors for GLMs

- Suppose we have a logistic regression model with likelihood

$$\mathcal{L}(\boldsymbol{\beta}|D) \propto \exp \left\{ \sum_{i=1}^n \left( y_i \mathbf{x}'_i \boldsymbol{\beta} - \log \left( 1 + e^{\mathbf{x}'_i \boldsymbol{\beta}} \right) \right) \right\}.$$

- The Fisher information matrix is given by

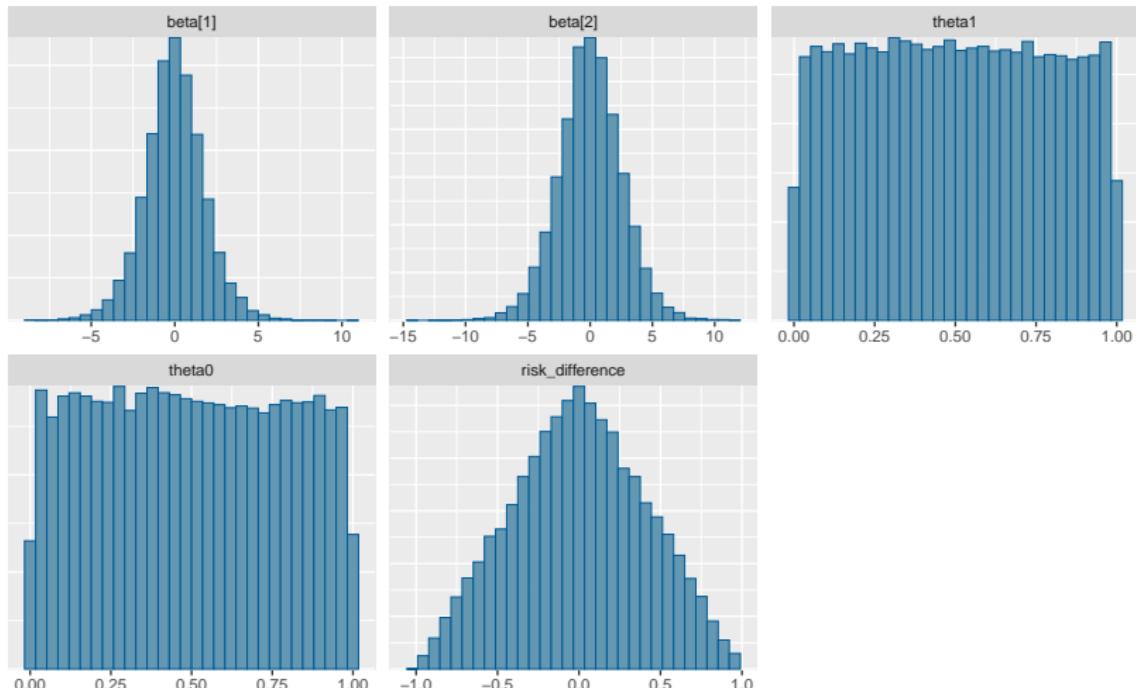
$$\mathcal{I}(\boldsymbol{\beta}) = E \left[ -\frac{\partial^2 \log \mathcal{L}}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} \right] = \mathbf{X}' \mathbf{W}(\boldsymbol{\beta}) \mathbf{X},$$

where  $\mathbf{W} = \text{diag}\{p_i(1-p_i), i = 1, \dots, n\}$ .

- It follows that Jeffreys' prior is given by

$$\pi_J(\boldsymbol{\beta}) \propto |\mathbf{X}' \mathbf{W}(\boldsymbol{\beta}) \mathbf{X}|^{1/2}.$$

# Jeffreys' prior for GLMs



Jeffreys' prior for a logistic regression model with intercept and treatment effect only.

## Other noninformative priors for regression

- Despite the attractive features of Jeffreys' prior, it is not often used in practice.
- Note that for survival models, Jeffreys' prior requires a model for the **censoring distribution**, which we are not often willing to make.
- Others have tried to use the **observed** Fisher information matrix, defined as

$$\mathcal{J}(\boldsymbol{\theta}) = -\frac{1}{n} \sum_{i=1}^n \frac{\partial^2 \log f(y_i | \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'},$$

but this **depends on the observed data**  $\mathbf{y} = (y_1, \dots, y_n)'$ , so this really isn't a prior.

## Other noninformative priors for regression

- In practice, we typically elicit

$$\pi(\beta) \propto 1 \quad \text{or} \quad \beta \sim N_p(\mathbf{0}, \sigma_0^2 \mathbf{I}_p),$$

where  $\sigma_0$  is large (e.g.,  $\sigma_0 = 10$ ).

- These priors will yield similar results, but note that only the latter guarantees posterior propriety for all models.
- This latter prior is closely related to **ridge regression**, e.g., the log posterior density is

$$\log p(\beta | \mathbf{y}) \propto \log L(\beta | \mathbf{y}) + \sigma_0^2 \sum_{j=1}^p \beta_j^2.$$

- In fact, **any** “penalized likelihood” approach may be interpreted as the **posterior mode** under a certain prior.

## Information matrix priors

- For LMs and GLMs, Jeffreys' prior requires the computation of determinants.
- As a result, these priors are only feasible for  $n > p$  models.
- For high dimensions, we need an alternative noninformative prior.
- The [information matrix](#) (IM) prior of Ibrahim and Gupta (2009) [4] can be thought of as a generalization of Jeffreys' prior.

## Information matrix priors

- For regression models with  $p < n$ , the IM prior is given as

$$\pi_{\text{IM}}(\boldsymbol{\beta} | \boldsymbol{\mu}_0, c_0) \propto |\mathcal{I}(\boldsymbol{\beta})|^{1/2} \exp \left\{ -\frac{1}{2c_0} (\boldsymbol{\beta} - \boldsymbol{\mu}_0)' \mathcal{I}(\boldsymbol{\beta}) (\boldsymbol{\beta} - \boldsymbol{\mu}_0) \right\}.$$

- For the normal linear model with unknown variance,  $\mathcal{I}(\boldsymbol{\beta}) = \frac{\mathbf{X}'\mathbf{X}}{\sigma^2}$ , so the IM prior is just a multivariate normal prior

$$\pi_{\text{IM}}(\boldsymbol{\beta} | \sigma^2, \boldsymbol{\mu}_0, c_0) = N(\boldsymbol{\beta} | \boldsymbol{\mu}_0, c_0 \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}).$$

- The **location parameter**  $\boldsymbol{\mu}_0$  may be set as **0**.
- The **dispersion parameter**  $c_0 > 0$  controls how informative the prior is.
- Note that, as  $c_0 \rightarrow \infty$ ,  $\pi_{\text{IM}}(\boldsymbol{\beta} | \boldsymbol{\mu}_0, c_0) \rightarrow$  Jeffreys' prior.

## Information matrix priors

- If  $p > n$ , we can no longer use the IM prior even when  $\sigma_0^2 > 0$  since the Fisher information matrix is singular (and, hence, the determinant of it is zero).
- We can introduce a **ridge parameter**,  $\lambda_0 > 0$ , to ensure positive definiteness. This results in the **information matrix ridge** (IMR) prior

$$\begin{aligned}\pi_{\text{IMR}}(\boldsymbol{\beta} | \boldsymbol{\mu}_0, c_0) &\propto |\mathcal{I}(\boldsymbol{\beta}) + \lambda_0 \mathbf{I}_p|^{1/2} \\ &\exp \left\{ -\frac{1}{2c_0\sigma^2} (\boldsymbol{\beta} - \boldsymbol{\mu}_0)' [\mathcal{I}(\boldsymbol{\beta}) + \lambda_0 \mathbf{I}_p] (\boldsymbol{\beta} - \boldsymbol{\mu}_0) \right\}\end{aligned}$$

- For example, for the normal linear model, the IMR prior is a normal prior given by

$$\boldsymbol{\beta} \sim N_p(\boldsymbol{\mu}_0, c_0\sigma^2 [\mathbf{X}'\mathbf{X} + \lambda_0 \mathbf{I}_p]^{-1})$$

## Markov chain Monte Carlo (MCMC)

- Thus far, we have only been able to derive posterior densities when the prior is conjugate.
- In most cases, we will not use conjugate priors either because (1) they do not exist; or (2) they do not lead to analytically tractable marginal likelihoods.
- For example, Jeffreys' prior for the logistic regression model does not lead to an analytically tractable posterior (nor prior!) density.
- When the posterior is not analytically tractable, we turn to MCMC methods for posterior inference.
- We focus less on MCMC algorithms and more on implementation and diagnostics through the useful R packages `cmdstanr`, `posterior`, and `bayesplot`.

## Stan basics I

- stan is a programming language that implements a gradient-based MCMC algorithm called Hamiltonian Monte Carlo (HMC).
- On the backend, Stan code is converted to C++ code.
- The primary limitation of Stan is that it cannot sample discrete parameters (but they can be integrated out of the model and generated later).

## Stan basics II

- Required blocks:
  - ① `data{}` - contains data and (possibly) hyperparameters.
  - ② `parameters{}` - contains names of parameters in the model and their types (e.g., `real`, `vector`).
  - ③ `model{}` - contains the likelihood and prior.
- Optional blocks:
  - ① `functions{}` - self explanatory.
  - ② `transformed data{}` - self explanatory.
  - ③ `transformed parameters{}` - transformation of parameters to be used **within** the model block.
  - ④ `generated quantities{}` - transformation of parameters to be done **after** all sampling (unused in model block).

## Stan: Bernoulli proportion example

- The below code implements a Bernoulli proportion with a  $U(0, 1)$  prior.

```
data {  
    int<lower=0> n;                      // number of observations  
    array[n] int<lower=0,upper=1> y; // integer array of size n giving Bernoulli responses  
}  
parameters {  
    real<lower=0,upper=1> p;    // Bernoulli proportion  
}  
model {  
    // prior  
    p ~ uniform(0, 1);    // not necessary since bounded parameters default to uniform prior  
    // likelihood  
    for ( i in 1:n )      // avoid looping whenever possible—see next slide  
        y[i] ~ bernoulli(p)  
}
```

## Stan: Equivalent (and faster) code

- Some distributions allow for vectorization. The same code can be written more efficiently as follows:

```
data {
  int<lower=0> n;                                // number of observations
  array [n] int<lower=0,upper=1> y; // integer array of size n giving Bernoulli responses
}
parameters {
  real<lower=0,upper=1> p;    // Bernoulli proportion
}
model {
  // likelihood
  y ~ bernoulli(p);
}
```

# Stan: Implementation of DY Prior ( $\mu$ -scale)

- Recall that for Bernoulli models, the DY prior on the  $\mu$ -scale is a beta prior, i.e.,

$$\pi_{\text{DY}}(\mu | n_0, \bar{y}_0) \propto \mu^{n_0 \bar{y}_0 - 1} (1 - \mu)^{n_0(1 - \bar{y}_0) - 1}.$$

```
data {
  int<lower=0> n;                                // number of observations
  array[n] int<lower=0,upper=1> y; // integer array of size n giving Bernoulli responses
  real<lower=0> n0;                            // prior sample size
  real<lower=0,upper=1> ybar0;      // prior prediction
}
transformed data {
  real shape1 = n0 * ybar0;
  real shape2 = n0 * (1 - ybar0);
}
parameters {
  real<lower=0,upper=1> p;    // Bernoulli proportion
}
model {
  // DY prior is a beta prior; we could do shape calculations here, but that wastes time
  p ~ beta(shape1, shape2);
  // likelihood
  y ~ bernoulli(p);
}
```

# Stan: Implementation of DY Prior ( $\theta$ -scale)

- Recall that for Bernoulli models, the posterior under the DY prior for a Bernoulli model is given by

$$\pi_{\text{DY}}(\theta | n_0, \bar{y}_0) \propto \exp \left\{ (n + n_0) \left[ \theta \left( \frac{n\bar{y} + n_0\bar{y}_0}{n + n_0} \right) - \log [1 + \exp(\theta)] \right] \right\}$$

```
data {  
    int<lower=0> n;                      // number of observations  
    array[n] int<lower=0,upper=1> y; // integer array of size n giving Bernoulli responses  
    real<lower=0> n0;                  // prior sample size  
    real<lower=0,upper=1> ybar0;      // prior prediction  
}  
transformed data {  
    real ybar = mean(y);  
    real post_precision = n + n0;  
    real post_shape = (n * ybar + n0 * ybar0) * inv(post_precision);  
}  
parameters {  
    real theta; // theta = log(p / (1 - p)) is unbounded  
}  
model {  
    // Increase log target density by DY posterior:  
    target += post_precision * (theta * post_shape + log1p_exp(theta));  
}  
generated quantities {  
    real p = inv_logit(theta); // could be in transformed parameters{}, but would be slower  
}
```

## Stan: A warning on parameter transformations

- It is advised to declare the parameter that will be used in `model` in the `parameters{}`.
- Example: the following would result in the wrong implementation of the DY prior:

```
data {  
    int<lower=0> n;                      // number of observations  
    array[n] int<lower=0,upper=1> y; // integer array of size n giving Bernoulli responses  
    real<lower=0> n0;                  // prior sample size  
    real<lower=0,upper=1> ybar0;     // prior prediction  
}  
parameters {  
    real<lower=0,upper=1> p; // bernoulli success probability  
}  
transformed parameters {  
    real theta = logit(p);  
}  
model {  
    // DY prior on theta scale—results in INCORRECT posterior  
    target += n0 * (theta * ybar0 - log1p_exp(theta));  
    y ~ bernoulli(p);  
}
```

- This can be fixed by adding `log(abs(jacobian))` to the log target density in the `model` block

```
target += -log(p) - log(1 - p);
```

## Part 2: Informative Prior Elicitation

## Informative priors

## Informative priors

- Loosely speaking, **informative priors** are priors that have an impact on the posterior density.
- There are various characterizations of informativeness.
- Two important criteria are:
  - ① Prior probability of a claim.
  - ② Effective sample size (ESS).

## Prior probability of a claim

- In clinical trials, it is common to use **posterior probabilities** to ascertain whether a trial is successful.
- For example, if  $\Delta$  is a treatment effect (e.g., a hazard ratio), we declare the trial to be successful if

$$\Pr(\Delta < 1|D) > 1 - \alpha/2.$$

- The **prior probability of a claim** (PPC) is simply the *prior probability* of the success criterion, i.e., for a prior  $\pi(\Delta)$ ,

$$\text{PPC} = \Pr_{\pi}(\Delta < 1) = \int_{-\infty}^1 \pi(\Delta) d\Delta$$

## Effective sample size

- A second useful metric is **effective sample size** (ESS).
- The ESS quantifies how many extra patients the prior was worth.
- For example, consider  $y|\theta \sim \text{Bin}(n, p)$  and  $\theta \sim \text{Beta}(\alpha, \beta)$ . Then  $\theta|y \sim \text{Beta}(y + \alpha, n - y + \beta)$ .
- This posterior is equivalent to that of

$$\tilde{p}(\theta|y) \propto \mathcal{L}(\theta|y + \alpha, n - y + \beta).$$

- Hence, the ESS is  $n + \alpha + \beta$  (i.e., the prior was worth  $\alpha + \beta$  patients).

## Effective sample size

- Under non-conjugate settings, it is not possible to analytically obtain the ESS.
- While many metrics have been proposed, among the most simple is Pennello and Thompson (2008) [5].
- Their measure is

$$\text{ESS} = n \times \frac{\text{Var}(\theta|D, \text{noninformative prior})}{\text{Var}(\theta|D, \text{informative prior})}$$

- For example, suppose  $y_i \sim N(\mu, 1)$  and  $\mu \sim N(\mu_0, \tau_0^{-1})$ . Then  $\mu|\mathbf{y} \sim N\left(\frac{n\bar{y} + \tau_0\mu_0}{n + \tau_0}, [n + \tau_0]^{-1}\right)$  and hence the posterior ESS is given by

$$\text{ESS} = n \times \frac{(1/n)}{1/(n + \tau_0)} = n + \tau_0$$

## Effective sample size

- For non-conjugate priors, the following algorithm can be used to compute the Pennello and Thompson ESS:

- 1 Sample from the posterior under a noninformative prior.
- 2 Sample from the posterior under an informative prior.
- 3 Compute ESS via

$$\text{ESS} = n \times \frac{\text{Var}(\theta|D, \text{noninformative prior})}{\text{Var}(\theta|D, \text{informative prior})}.$$

- Note that the Pennello and Thompson approach assumes that the precision (i.e., inverse variance) is proportional to the sample size, which is true asymptotically by the **Bayesian CLT**, which we now review.

## Bayesian CLT

- Like the Frequentist paradigm, there is a central limit theorem (CLT) for the Bayesian paradigm.
- Suppose that  $\pi(\theta)$  is a prior for  $\theta \in \Theta$ , where  $\Theta$  is the parameter space for  $\theta$ .
- Suppose further that  $\pi(\theta) > 0$  for all  $\theta \in \Theta$ .
- The Bayesian CLT states

$$p(\theta|y) \rightarrow N_p \left( \theta \left| \hat{\theta}, [H(\hat{\theta})]^{-1} \right. \right)$$

as  $n \rightarrow \infty$ , where

①  $\hat{\theta}$  = posterior mode

②  $H(\cdot) = -\text{Hessian matrix of log posterior density} = -\frac{\partial^2 \log p(\theta|y)}{\partial \theta \partial \theta'}$

## Bayesian CLT

- An important implication of the Bayesian CLT is that, in large samples, the prior does not really matter.
- Suppose that  $\theta^*$  is the true parameter. As long as  $\pi(\theta^*) > 0$ , frequentist operating characteristics (e.g., type I error rates) can be recovered asymptotically.
- However, in the presence of a **prior-data conflict**, a substantially larger sample size will be required to obtain type I error rates close to their nominal value.
- Still, the Bayesian CLT bridges the gap between the frequentist and Bayesian paradigms, and provides a small amount of protection against prior-data conflicts.

Informative prior elicitation based on historical data

## Historical data

- In clinical trials, it is common to possess historical data.
- There are various examples:
  - ① Phase II study data could be used in a Phase III study.
  - ② Adult study data could be used in a pediatric study.
- Why use historical data?
  - ① Ideally, you wouldn't have to.
  - ② Rare diseases
  - ③ Pediatric studies
  - ④ Impractical / too expensive
- In this section, we will study informative prior elicitation, mostly on the basis of historical data.

## The Power Prior

- The power prior of Ibrahim and Chen (2000) [6] (reviewed in Ibrahim et al (2015) [7]) has emerged as a useful class of informative priors for a variety of situations in which historical data is available.
- Suppose we have **historical data**  $D_0 = (n_0, \mathbf{y}_0, \mathbf{X}_0)$  from a previous study.
  - ▶  $n_0$  = sample size of the historical data.
  - ▶  $\mathbf{y}_0$  = response vector of historical data.
  - ▶  $\mathbf{X}_0 = n_0 \times p$  design matrix of historical data.
- Let  $\boldsymbol{\theta}$  denote a parameter vector, e.g.,  $\boldsymbol{\theta} = (\boldsymbol{\beta}', \sigma^2)'$  for a normal linear model.
- Further, let  $\pi_0(\boldsymbol{\theta})$  denote the prior distribution for  $\boldsymbol{\theta}$  before the historical data  $D_0$  is observed. The **initial prior** distribution for  $\boldsymbol{\theta}$ .
- Typically,  $\pi_0(\boldsymbol{\theta})$  is taken to be noninformative, and oftentimes improper.

## The Power Prior

- Let the data from the *current* study be denoted by  $D = (n, \mathbf{y}, \mathbf{X})$ .
- Further, denote the likelihood for the current study by  $\mathcal{L}(\theta|D)$ , where  $\theta$  is a vector of indexing parameters.
- Thus,  $\mathcal{L}(\theta|D)$  is a general likelihood function for an arbitrary parametric model, such as linear models, generalized linear model, random effects model, nonlinear model, or a survival model with censored data.

## The Power Prior

- Given the discounting parameter  $a_0 \in [0, 1]$ , we define the **power prior** distribution of  $\theta$  for the current study as

$$\pi_{\text{PP}}(\theta|D_0, a_0) \propto \mathcal{L}(\theta|D_0)^{a_0} \pi_0(\theta), \quad (2)$$

where  $a_0$  weights the historical data relative to the likelihood of the current study,

- Thus the hyperparameter  $a_0$  controls the influence of the historical data relative to  $\mathcal{L}(\theta|D)$ .
- The hyperparameter  $a_0$  can be interpreted as a precision parameter for the historical data and  $a_0 n_0$  as effective sample size.
- In general, we restrict  $a_0 n_0 \leq n$  so that the historical data does not have more weight on the posterior density than the current data.

## The Power Prior

- Thus it is scientifically more sound to restrict the range of  $a_0$  to be between 0 and 1, and thus we take  $0 \leq a_0 \leq 1$ .
- Setting  $a_0 = 1$ , the power prior corresponds to the update of  $\pi_0(\theta)$  using Bayes theorem. That is, it corresponds to the posterior distribution of  $\theta$  based on the historical data.
- When  $a_0 = 0$ , then the prior does not depend on the historical data  $D_0$ ; in this case,  $\pi(\theta|D_0, a_0 = 0) \equiv \pi_0(\theta)$ .

## The Power Prior

- The parameter  $a_0$  allows the investigator to control the influence of the historical data on the analysis of the current data.
- Such control is important in cases where there is heterogeneity between the previous and current studies, or when the sample sizes of the two studies are quite different.
- In practice, one may prespecify  $a_0$  or try to determine a reasonable value based on observed data using a measure of prior data conflict (see for example Ibrahim et al, 2015 [7]).

# The Power Prior

- Since the power prior is basically a likelihood function raised to a power, it shares all of the properties that likelihood functions have, and therefore has many nice properties.
  - ▶ **Propriety:** techniques for showing propriety of  $\pi_{\text{PP}}(\theta|D_0, a_0)$  are the same as those for showing propriety for a posterior distribution based on a dataset  $D_0$  with likelihood function  $\mathcal{L}(\theta|D_0)$ , and prior  $\pi_0(\theta)$ .
  - ▶ **Computation:** Gibbs sampling and related MCMC techniques are the same as those of a posterior distribution based on a dataset  $D_0$  with likelihood function  $\mathcal{L}(\theta|D_0)$ , and prior  $\pi_0(\theta)$ .
  - ▶ **Asymptotics:** The asymptotic distribution of the power prior is a normal distribution.

## The Power Prior

- For the asymptotic property, it can be shown that as  $n_0 \rightarrow \infty$ ,

$$\pi(\theta|D_0, a_0) \approx \phi_p(\theta|\hat{\theta}_0, a_0^{-1}H_0^{-1}(\hat{\theta}_0)), \quad (3)$$

where  $\hat{\theta}_0$  is the mode of the power prior and

$$H_0(\theta) = -\frac{\partial^2 \log \pi(\theta|D_0, a_0)}{\partial \theta \partial \theta'}$$

is the negative Hessian matrix evaluated at  $\theta$ .

- The posterior  $p(\theta|D, D_0, a_0)$  has a similar approximation where  $H(\theta)$  is based on the combined log-likelihoods using **case weights** and  $\hat{\theta}$  is the posterior mode (Psioda and Ibrahim, 2019) [8].
- This approximation is very useful for simulation-based SSD as it allows one to avoid the need for MCMC methods which can be very time consuming.

## Example: Analysis of the PLUTO trial

- Based on data generated by two well-controlled pivotal trials in adult systemic lupus erythematosus (SLE), the BLISS-52 and BLISS-76 trials, the FDA approved belimumab for the treatment of adults with active, seropositive SLE (who are already on SOC).

		Study 1056			Study 1057	
	Placebo N=275	Belimumab 1 mg/kg N=271	Belimumab 10 mg/kg N=273	Placebo N=287	Belimumab 1 mg/kg N=288	Belimumab 10 mg/kg N=290
<b>Response, n (%)</b>	93 (34)	110 (41)	118 (43)	125 (44)	148 (51)	167 (58)
<b>Observed difference</b>	-	7%	9%	-	8%	14%
<b>Odds ratio (95% CI)</b>	-	1.3 (0.9, 1.9)	1.5 (1.1, 2.1)	-	1.6 (1.1, 2.2)	1.8 (1.3, 2.6)

Source: Review of BLA 125370 Belimumab IV dated February 18, 2011.

BLISS Trials Primary Endpoint Data: SRI Response Rate

## Example: Analysis of the PLUTO trial

The Pediatric Lupus Trial of Belimumab Plus Background Standard Therapy (PLUTO) is a Phase II, multicenter, randomized, double-blind trial evaluating the efficacy, safety, and pharmacokinetics of IV belimumab versus placebo plus SOC in childhood-onset Systemic Lupus Erythematosus patients aged 5–17 years (NCT01649765).

At 52 weeks, clinical response was observed in  $y_1 = 28$  of  $n_1 = 53$  treated patients and  $y_0 = 17$  out of  $n_0 = 39$  placebo patients.

- $y_1 = 28$ ;  $n_1 = 53$  and  $y_0 = 17$ ;  $n_0 = 39$ .
- $\bar{y}_1 - \bar{y}_0 = 0.092$ .

## Example: Power prior

- Let  $\delta(\theta) = \theta_1 - \theta_0$  be the difference in response probabilities for participants treated with 10 mg/kg belimumab ( $\theta_1$ ) compared to placebo ( $\theta_0$ ).
- We formulate a power prior based on  $\theta_1$  and  $\theta_0$  noting that we may compute the posterior for any function of them using MCMC, namely  $\delta(\theta) = \theta_1 - \theta_0$ .

## Example: Power prior

- Specifically, we have

$$\pi(\boldsymbol{\theta}|D_0, \mathbf{a}_0) \propto \mathcal{L}(\boldsymbol{\theta}|D_{01})^{a_{01}} \mathcal{L}(\boldsymbol{\theta}|D_{02})^{a_{02}} \pi_0(\boldsymbol{\theta}), \quad (4)$$

where

- ▶  $D_{01} = \{(y_{01}, n_{01}) = (93, 275), (y_{11}, n_{11}) = (118, 273)\}$ ,
- ▶  $D_{02} = \{(y_{02}, n_{02}) = (125, 287), (y_{12}, n_{12}) = (167, 290)\}$ ,
- ▶ The likelihood for the  $s^{th}$  data set may be written as

$$\mathcal{L}(\boldsymbol{\theta}|D_{0s}) = \theta_0^{y_{0s}} (1 - \theta_0)^{n_{0s} - y_{0s}} \theta_1^{y_{1s}} (1 - \theta_1)^{n_{1s} - y_{1s}},$$

- ▶  $\pi_0(\boldsymbol{\theta}) = \pi_0(\theta_0)\pi_0(\theta_1)$  with  $\pi_0(\theta_j) = \text{Beta}(\theta_j|\alpha_0, \beta_0)$ , and
- ▶  $a_{0s} \in [0, 1]$ .

## Power Prior: Stan implementation

- R and Stan code (in rmarkdown) to implement the power prior for this example is available at <https://github.com/ethan-alt/IntroBayesianAnalysis/tree/main/Examples>
- Click here to see a summary of the data analysis.

# The Normalized Power Prior

- Alternatively, one may treat  $a_0$  as a random variable and assign it a prior distribution. With appropriate normalization, this results in a **normalized power prior** (Duan et al, 2006 [9]).
- The normalized power prior is given by

$$\begin{aligned}\pi(\theta, a_0 | D_0) &= \frac{1}{c(a_0, D_0)} \mathcal{L}(\theta | D_0)^{a_0} \pi_0(\theta) \pi_0(a_0), \\ &= \pi(\theta | D_0, a_0) \pi_0(a_0)\end{aligned}\quad (5)$$

where

- ▶  $c(a_0, D_0) = \int \mathcal{L}(\theta | D_0)^{a_0} \pi_0(\theta) d\theta$ , and
- ▶  $\pi_0(a_0)$  is an initial prior for  $a_0$  often taken to be a beta distribution.
- Often the integral  $c(a_0, D_0)$  has no closed-form. High-quality approximation methods have been developed for efficient MCMC sampling (Carvalho and Ibrahim, 2021 [10]).

## Normalized power prior: PLUTO trial

- The power prior with an initial beta prior is given by

$$\begin{aligned}\pi_{\text{PP}}(\boldsymbol{\theta} | \mathcal{D}_0, \mathbf{a}_0) &\propto \prod_{j=0}^1 \left\{ \theta_j^{\alpha_{0j}-1} (1-\theta_j)^{\beta_{0j}-1} \prod_{s=1}^2 \theta_j^{a_{0s}y_{0sj}} (1-\theta_j)^{a_{0s}(n_{0sj}-y_{0sj})} \right\} \\ &= \prod_{j=0}^1 \theta_j^{\alpha_{0j}^*-1} (1-\theta_j)^{\beta_{0j}^*-1},\end{aligned}$$

where

- $\alpha_{0j}^* = \sum_{s=1}^2 a_{0s} y_{0sj} + \alpha_{0j}$ ,
- $\beta_{0j}^* = \sum_{s=1}^2 a_{0s} (n_{0sj} - y_{0sj}) + \beta_{0j}$ .

- The kernel of the power prior is the kernel of a  $\text{Beta}(\alpha_{0j}^*, \beta_{0j}^*)$  density, so the normalizing constant is  $c(a_0, D_0) = B(\alpha_{0j}^*, \beta_{0j}^*)$ .
- We elicit a  $\text{Beta}(\gamma_{0j}, \eta_{0j})$  prior on  $a_0$ .

## Normalized power prior: PLUTO trial

- The power prior with an initial beta prior is given by

$$\begin{aligned}\pi_{\text{PP}}(\boldsymbol{\theta} | \mathbf{D}_0, \mathbf{a}_0) &\propto \prod_{j=0}^1 \left\{ \theta_j^{\alpha_{0j}-1} (1-\theta_j)^{\beta_{0j}-1} \prod_{s=1}^2 \theta_j^{a_{0s}y_{0sj}} (1-\theta_j)^{a_{0s}(n_{0sj}-y_{0sj})} \right\} \\ &= \prod_{j=0}^1 \theta_j^{\alpha_{0j}^*-1} (1-\theta_j)^{\beta_{0j}^*-1},\end{aligned}$$

where

- $\alpha_{0j}^* = \sum_{s=1}^2 a_{0s} y_{0sj} + \alpha_{0j}$
- $\beta_{0j}^* = \sum_{s=1}^2 a_{0s} (n_{0sj} - y_{0sj}) + \beta_{0j}$
- The kernel of the power prior is the kernel of a  $\text{Beta}(\alpha_{0j}^*, \beta_{0j}^*)$  density, so the normalizing constant is  $c(a_0, D_0) = B(\alpha_{0j}^*, \beta_{0j}^*)$

## Normalized power prior: PLUTO trial

- R and Stan code (in `rmarkdown`) to implement the normalized power prior for this example is available at  
[https://github.com/ethan-alt/IntroBayesianAnalysis/blob/main/Examples/PLUTO\\_normalizedPower.Rmd](https://github.com/ethan-alt/IntroBayesianAnalysis/blob/main/Examples/PLUTO_normalizedPower.Rmd)
- Click [here](#) to see a summary of the data analysis.

## Bayesian hierarchical model (BHM)

- For the BHM, we consider a logistic regression model parametrization given by

$$\text{logit} [P(Y_{hi} = 1 | z_{hi})] = \beta_{0h} + \beta_{1h} z_{hi},$$

for participant  $i = 1, \dots, n_h$  from study  $h = 1, 2, 3$ .

- Here  $z_{hi}$  is an indicator for whether participant  $i$  was randomized to receive belimumab.
- We consider  $h = 1$  to correspond to the PLUTO (pediatric) data and  $h = 2$  and  $h = 3$  to correspond to the BLISS-52 and BLISS-76 (adult) data, respectively.

## Bayesian hierarchical model

- The hierarchical prior for the intercepts is given as follows.

$$\beta_{0h} | \beta_0, \tau_0 \sim \text{Normal}(\beta_0, \text{sd} = \tau_0)$$

$$\beta_0 \sim \text{Normal}(0, \text{sd} = 3)$$

$$\tau_0 \sim \text{Half-Cauchy}(0, 1)$$

- The hierarchical prior for the treatment effects is given as follows.

$$\beta_{1h} | \beta_1, \tau_1 \sim \text{Normal}(\beta_1, \text{sd} = \tau_1)$$

$$\beta_1 \sim \text{Normal}(0, \text{sd} = 3)$$

$$\tau_1 \sim \text{Half-Cauchy}(0, 1)$$

- The non-conjugate hyper-priors for  $\tau_0$  and  $\tau_1$  are proper, but noninformative ( $\text{Var}(\tau_j) = \infty$ ).
- The Half-Cauchy prior was suggested by Gelman [11].

## Bayesian hierarchical model

- The hierarchical parameters  $\tau_0$  and  $\tau_1$  quantify the degree to which the three data sets agree (**between-study heterogeneity**).
- For example, if  $\Pr(\tau_1 < 1/3|\mathbf{y})$  is large, then the treatment effects are likely to be within 1 standard deviation of each other.
- Special choices of  $a_0$  in the power prior and particular priors in the BHM yield equivalent results [12].
- In general, it is difficult to ascertain the level of borrowing of the prior.
- The choice of priors must be justified to regulators, which can sometimes be difficult.

## Bayesian hierarchical model: PLUTO Example

- R and Stan code for the implementation of the BHM for the PLUTO study is available at  
[PLUTO.bhm.Rmd](#)
- Click here for the data analysis results.

## Meta-analytic predictive (MAP) priors

- Note that the BHM, which is also referred to as a **meta-analytic combined** (MAC) approach, is problematic for design since we must possess the current data.
- The **meta-analytic predictive** (MAP) prior is the prior for  $(\beta_0, \beta_1)$  induced by the BHM.
- It can be thought of as a multi-step approach:
  - ▶ Estimate the posterior distribution of the historical data sets using a BHM, i.e.,  $p(\beta_{0h}, \beta_{1h}, \beta_0, \beta_1, \tau_0, \tau_1 | \mathbf{D}_0)$  for  $h \geq 2$  (using  $h = 1$  as current data).
  - ▶ The prior for  $(\beta_{01}, \beta_{11})$  is obtained from the **predictive distribution**, i.e.,

$$\pi_{\text{MAP}}(\beta_{01}, \beta_{11} | \mathbf{D}_0) = \int \phi(\beta_{01} | \beta_0, \tau_0) \phi(\beta_{11} | \beta_1, \tau_1) p(\beta_0, \beta_1, \tau_0, \tau_1 | \mathbf{D}_0) d\boldsymbol{\eta},$$

where  $\boldsymbol{\eta} = (\beta_0, \beta_1, \tau_0, \tau_1)'$  are the hierarchical parameters.

## Meta-analytic predictive (MAP) priors

- In general,  $\pi_{\text{MAP}}$  will not have a closed form, although it is easy to sample from it.
- It has been recommended to use mixtures of distributions to obtain an accurate approximation of  $\pi_{\text{MAP}}$ , e.g.,

$$\pi_{\text{MAP}}(\beta_{01}, \beta_{11} | \mathbf{D}_0) = \sum_{k=1}^K \pi_k f_k(\beta_{01}, \beta_{11} | \gamma_k)$$

- We can treat samples from  $\pi_{\text{MAP}}$  as data and use any mixture model fitting software.
- This method can be particularly attractive when the  $f_k$ 's are conjugate for computation and interpretation.
- Note that the MAP allows for characterization of the level of informativeness of the BHM.

## Mixture priors

- Mixture priors combine one or more informative priors with a vague prior via a **mixture distribution**:

$$\pi(\theta) = \sum_{i=1}^I \gamma_{li} f_{li}(\theta | \eta_i) + \gamma_V f_V(\theta | \eta_V), \quad \sum_{i=1}^I \gamma_{li} + \gamma_V = 1.$$

- Note that each  $f_{li}$  and  $f_V$  must be a **properly normalized density**.
- In the context of historical data, the  $f_{li}$ 's may be chosen as a normal approximation to the posterior of  $\theta$  for each historical data set.
- The vague prior could be, e.g.,  $\theta \sim N(0, 10^2)$  for  $\theta \in (-\infty, \infty)$  or a uniform prior if  $\theta$  is bounded.

## Mixture priors

- The posterior density under a mixture prior is also a mixture with **updated weights**.
- E.g., for a Bernoulli model, the posterior is given by

$$\begin{aligned} p(\theta|\mathbf{y}) &\propto \theta^{n\bar{y}}(1-\theta)^{n(1-\bar{y})} \sum_{k=1}^K \gamma_k \frac{\theta^{\alpha_k-1}(1-\theta)^{\beta_k-1}}{B(\alpha_k, \beta_k)}, \\ &\propto \sum_{k=1}^K \gamma_k \frac{B(n\bar{y} + \alpha_k, n(1 - \bar{y}) + \beta_k)}{B(\alpha_k, \beta_k)} \frac{\theta^{n\bar{y}+\alpha_k-1}(1-\theta)^{n(1-\bar{y})+\beta_k-1}}{B(n\bar{y} + \alpha_k, n(1 - \bar{y}) + \beta_k)}, \\ &\propto \sum_{k=1}^K \tilde{\gamma}_k f_\beta(\theta|n\bar{y} + \alpha_k, n(1 - \bar{y}) + \beta_k), \end{aligned} \tag{6}$$

where

$$\tilde{\gamma}_k = \frac{\gamma_k B(n\bar{y} + \alpha_k, n(1 - \bar{y}) + \beta_k) / B(\alpha_k, \beta_k)}{\sum_{m=1}^K \gamma_m B(n\bar{y} + \alpha_m, n(1 - \bar{y}) + \beta_m) / B(\alpha_m, \beta_m)}$$

## Mixture priors

- It can be shown that the updated weights are a function of the **Bayes factor** comparing the fit of two different priors.
- The updated weights make mixture priors extremely attractive since resilience against prior-data conflict is built-in.
- The most challenging part of mixture priors is how to choose the weights.

## Mixture priors: PLUTO Example

- To obtain a robust mixture prior, we conduct separate analysis of the BLISS-52 and BLISS-76 trials via maximum likelihood:

Data set	Intercept	Treatment Effect
BLISS-52	-0.6714(0.1275)	0.3987 (0.1766)
BLISS-76	-0.2593(0.1190)	0.5651(0.1682)

MLE of historical data sets (SE in parentheses).

- We elicit a multivariate normal priors with mean = MLE, covariance = inverse Fisher information.

$$\blacktriangleright \text{BLISS-52: } f_{l1}(\beta | \mu_1, \Sigma_1), \Sigma_1 = \begin{pmatrix} 0.0162 & -0.0162 \\ -0.0162 & 0.0312 \end{pmatrix}$$

$$\blacktriangleright \text{BLISS-76: } f_{l2}(\beta | \mu_2, \Sigma_2), \Sigma_2 = \begin{pmatrix} 0.0142 & -0.0142 \\ -0.0142 & 0.0283 \end{pmatrix}$$

## Mixture priors: PLUTO Example

- For robustification, we mix these priors with a  $N_2(\mathbf{0}, 100^2 \mathbf{I}_2)$  vague prior.
- The implemented prior is thus:

$$\pi_{RM}(\beta|D_0) = \gamma_{I1} N_2(\beta|\mu_1, \Sigma_1) + \gamma_{I2} N_2(\beta|\mu_2, \Sigma_2) + \gamma_V N_2(\beta|\mathbf{0}, 10^2 \mathbf{I})$$

- We choose  $\gamma_{I1} = 0.3$ ,  $\gamma_{I2} = 0.3$ ,  $\gamma_V = 0.4$ .
- The code is available in [/Examples/PLUTO\\_robustMixture.qmd](#)
- Click [here](#) for the data analysis results .

## Robust MAP priors

- Recall that the MAP prior is the prior for  $\theta$  induced by the hierarchical model.
- The BHM has been criticized because it is difficult to control the level of borrowing.
- We have seen how Mixture priors can “robustify” informative priors via mixing with a vague prior.
- The **robust MAP** prior is a conglomeration of the two:

$$\pi_{\text{rMAP}}(\theta|\gamma, D_0) = \gamma\pi_{\text{MAP}}(\theta|D_0) + (1 - \gamma)\pi_v(\theta)$$

- The R package RBesT provides an implementation.

# Robust MAP priors

- A popular approach is to approximate  $\pi_{\text{MAP}}$  with a **finite mixture**.
- E.g., we may specify logit  $[\Pr(y_{0ji} = 1 | z_{0ji})] = \beta_{Hj0} + \beta_{Hj1}z_{0ji}$ ,
- We may elicit  $\beta_{Hjk} \sim N(\mu_{0k}, \tau_{0k}^{-1})$ ,  $j = 1, \dots, J$ ,  $k = 0, 1$ , applying the following algorithm for the approximation:
  - ① Obtain a sample  $\{(\beta_H^{(m)}, \mu_0^{(m)}, \tau_0^{(m)}, m = 1, \dots, M\}$  using the historical data sets via MCMC.
  - ② Draw  $\beta_k^{(m)} \sim N\left(\mu_{0k}^{(m)}, \left[\tau_{0k}^{(m)}\right]^{-1}\right)$ .
  - ③ Compute  $\theta_0^{(m)} = \text{logit}^{-1}(\beta_0^{(m)})$  and  $\theta_1^{(m)} = \text{logit}^{-1}(\beta_0^{(m)} + \beta_1^{(m)})$ .
  - ④ Approximate with mixture of beta priors

$$\hat{\pi}_{\text{MAP}}(\boldsymbol{\theta}) = \prod_{k=0}^1 \left\{ \sum_{l=1}^{L_k} [p_{kl} \times \text{Beta}(\theta_k | \alpha_{kl}, \beta_{kl})] \right\}$$

# Robust MAP priors

- Note that (4) on the previous slide is a mixture of conjugate priors.
- A mixture of conjugate priors is a conjugate prior, e.g.,

$$\begin{aligned} p(\theta_0|D) &\propto \theta_0^{y_0} (1-\theta_0)^{n_0-y_0} \times \left\{ \sum_{l=1}^{L_0} p_{0l} \times \frac{1}{B(\alpha_{0l}, \beta_{0l})} \times \theta_0^{\alpha_{0l}} (1-\theta_0)^{\beta_{0l}} \right\} \\ &= \sum_{l=1}^{L_0} \frac{p_{0l}}{B(\alpha_{0l}, \beta_{0l})} \times \theta_0^{y_0 + \alpha_{0l} - 1} (1-\theta_0)^{n_0 - y_0 + \beta_{0l} - 1} \\ &\propto \sum_{l=1}^{L_0} \frac{p_{0l} \times B(y_0 + \alpha_{0l}, n_0 - y_0 + \beta_{0l})}{B(\alpha_{0l}, \beta_{0l})} \times f_\beta(\theta_0 | y_0 + \alpha_{0l}, n_0 - y_0 + \beta_{0l}) \\ &\propto \sum_{l=1}^{L_0} \tilde{p}_{0l} \times f_\beta(\theta_0 | y_0 + \alpha_{0l}, n_0 - y_0 + \beta_{0l}), \end{aligned}$$

where

$$\tilde{p}_{0l} = \frac{\frac{\tilde{p}_{0l} \times B(y_0 + \alpha_{0l}, n_0 - y_0 + \beta_{0l})}{B(\alpha_{0l}, \beta_{0l})}}{\sum_{m=1}^{L_0} \frac{\tilde{p}_{0m} \times B(y_0 + \alpha_{0m}, n_0 - y_0 + \beta_{0m})}{B(\alpha_{0m}, \beta_{0m})}}$$

## Robust MAP priors

- Thus, a mixture of conjugate priors yields a conjugate posterior with updated mixture weights.
- It is worth mentioning that these approximations are attractive for **Bayesian design**.
- Most Bayesian design problems are **simulation-based**, where we simulate, say, 10,000 data sets and analyze the posterior of each.
- Since the posterior is a mixture of betas, we do not need to conduct MCMC sampling for inference. **This speeds things up greatly.**
- Note: the idea of approximating priors with conjugate priors is not specific to robust MAP priors. The approach is applicable to any prior.

## Robust MAP priors

- We implement the robust MAP prior for the PLUTO study.
- The R and Stan code is available at [Examples/PLUTO\\_rmap.Rmd](#)
- [Click here](#) for the data analysis results.

## Commensurate priors

- The commensurate prior (CP) (Hobbs et al, 2012) [13] is developed for single historical data set settings.
- The CP assumes

$$\pi_{\text{CP}}(\boldsymbol{\beta}, \boldsymbol{\beta}_0 | D_0) \propto \left[ \prod_{j=1}^J N_p(\beta_j | \beta_{0j}, \tau_j^{-1}) \right] \mathcal{L}(\boldsymbol{\beta}_0 | D_0) \pi_0(\boldsymbol{\beta}_0)$$

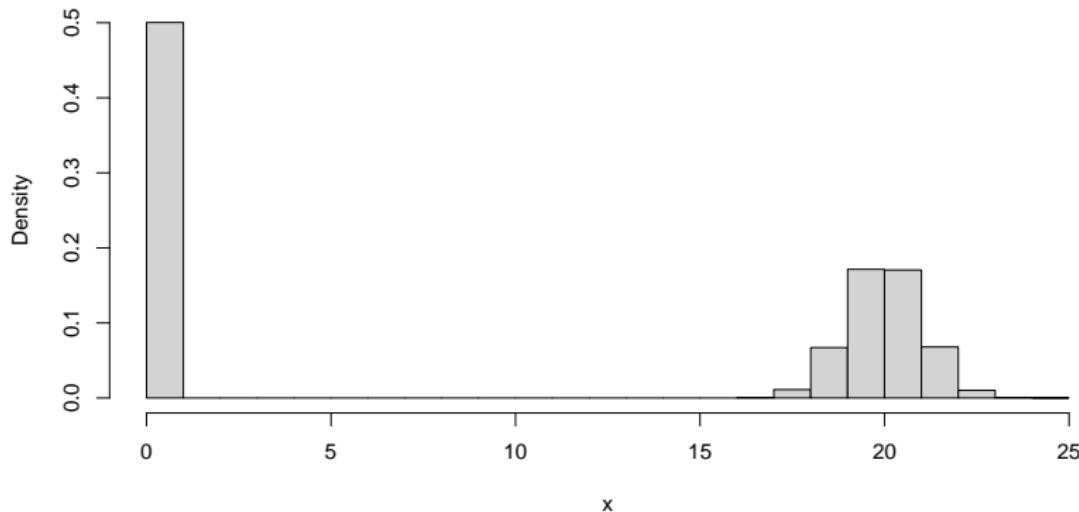
- Similar to the BHM, the CP does not assume that the regression coefficients are the same.
- The precision parameters  $\tau_j$ 's measure the **commensurability** of the current and historical data sets **for each parameter**.
- The authors recommend independent spike and slab priors on  $\tau_j$ 's, e.g.,

$$\pi(\tau_j) = \gamma N(\tau_j | 20, 1) + (1 - \gamma) U(\tau_j | 0.10, 0.50)$$

- $\gamma$  may be elicited or given a hyperprior.

## Commensurate priors

- We plot the marginal prior for  $\tau_j$  when  $\gamma \sim U(0, 1)$ .



Marginal prior for  $\tau_j$  when  $\tau_j|\gamma$  is spike and slab and  $\gamma \sim U(0, 1)$ .

## Commensurate priors: PLUTO Example

- We specify  $\gamma \sim U(0, 1)$ .
- The implementation is in `Examples/PLUTO_commensurate.Rmd`.
- Click [here](#) for the data analysis results.

## Historical Data: Advanced Topics

## Historical data: Advanced Topics

- Thus far, we have discussed informative prior elicitation mostly for i.i.d. models.
- In the sequel, we will focus on regression models.
- We motivate the following methods by the ESTEEM-I and ESTEEM-II trials for plaque psoriasis.

## The ESTEEM trials

- Double-masked, placebo controlled clinical trials evaluating the efficacy and safety of Apremilast, an oral phosphodiesterase 4 inhibitor, on patients with moderate-to-severe plaque psoriasis.
- Both studies conducted 2:1 randomization.
- Primary outcome: percentage of participants who achieved an improvement of  $\geq 75\%$  in the Psoriasis Area Severity Index (PASI) score.
- We will consider the percentage reduction in the PASI score as the primary outcome (as opposed to its dichotomization).
- NOTE: The data sets used are simulated based on the real data set.

# The ESTEEM trials

Characteristic	ESTEEM I, N = 844			ESTEEM II, N = 411		
	N	Placebo,	30 mg BID,	N	Placebo,	30 mg BID,
		N = 282 <sup>1</sup>	N = 562 <sup>1</sup>		N = 137 <sup>1</sup>	N = 274 <sup>1</sup>
Age	844	46.5 (12.7)	45.8 (13.1)	411	45.7 (13.4)	45.3 (13.1)
Smoker Category	844			411		
Current user		92 (33%)	202 (36%)		61 (45%)	101 (37%)
Not a current user		190 (67%)	360 (64%)		76 (55%)	173 (63%)
Prior use of Systemic Therapies	844			411		
N		132 (47%)	261 (46%)		64 (47%)	117 (43%)
Y		150 (53%)	301 (54%)		73 (53%)	157 (57%)
Baseline PASI Score	844	19.4 (7.4)	18.7 (7.2)	411	20.0 (8.0)	18.9 (7.1)
% Change in total PASI	837	-16.7 (31.5)	-52.1 (32.8)	405	-15.8 (41.3)	-50.9 (34.0)
Score at Wk 16						

Baseline characteristics of patients enrolled in the ESTEEM-I and ESTEEM-II trials. Continuous variables show mean (SD); categorical variables show N (%)

# The ESTEEM trials

- For ESTEEM-II (current data), we assume the linear model

$$y_i = \beta_0 + \beta_1 a_i + \mathbf{x}'_i \boldsymbol{\beta}_2 + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2).$$

- $y_i$  = percent change in PASI
- $a_i$  = treatment indicator (0 = placebo, 1 = Apremilast)
- $\mathbf{x}_i$  = vector of covariates
  - ① Age
  - ② Age<sup>2</sup>
  - ③ Smoking status (current user / not a current user)
  - ④ Prior use of systemic therapies (no / yes)
  - ⑤ Baseline PASI score

## Partial borrowing power prior

- Recall that the power prior for a parameter vector  $\theta = (\theta'_1, \theta'_2)$  is given by

$$\pi_{\text{PP}}(\theta_1, \theta_2 | D_0, a_0) \propto L(\theta_1, \theta_2 | D_0)^{a_0} \pi_0(\theta_1, \theta_2)$$

- Sometimes, we wish to only borrow information on  $\theta_1$ . This can happen in the contexts of
  - Borrowing from external controls.
  - Unavailable covariates.
  - Only wishing to borrow for the treatment effect.
- In these situations, we may use the partial borrowing power prior (PBPP) given by

$$\pi_{\text{PBPP}}(\theta_1) = \int \pi_{\text{PP}}(\theta_1, \theta_2 | D_0, a_0) d\theta_2$$

## Partial borrowing power prior

- At first glance, the PBPP seems unusable due to the possibly high-dimensional integral over  $\theta_2$ .
- Recall that we can always sample from a **univariate** distribution by sampling from a **multivariate** distribution and ignoring the margins we do not care about.
- E.g., we can sample from  $y_1 \sim N(\mu_1, \sigma_1^2)$  by sampling from

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \sim N \left( \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix} \right)$$

and ignoring the  $y_2$  samples.

- Applying the same logic, the sampling from the posterior under the PBPP is equivalent to sampling from

$$p(\theta_1, \theta_2 | D_0, a_0) \propto L(\theta_1 | D) L(\theta_1, \theta_2 | D_0)^{a_0} \pi_0(\theta_1, \theta_2)$$

and ignoring the  $\theta_2$  samples.

## Partial borrowing power prior: PLUTO example

- We implement the PBPP for the ESTEEM-II study, where we wish only to borrow from the ESTEEM-I controls.
- We will borrow information from the intercept and the covariate effects, but not the treatment effect.
- The code is available at [Examples/PLUTO\\_partialBorrowingPowerPrior.rmd](#) .
- [Click here for the data analysis results.](#)

## The scale-transformed power prior (straPP)

- Suppose that the current and historical data are related, but of different types, e.g.,
  - ①  $y_1 = \text{change in PASI score (continuous)} \rightarrow \theta$ .
  - ②  $y_0 = \text{reduction of } \geq 75\% \text{ of PASI score (binary)} \rightarrow \eta$ .
- While  $\theta, \eta$  are related, they are measured on different scales, violating the main assumption in the power prior (i.e,  $\theta = \eta$ ).
- The straPP (Alt et al. 2023) [14] is predicated on the assumption:

$$I_{\Theta}(\theta)^{1/2}\theta = I_{\boldsymbol{\eta}}(\boldsymbol{\eta})^{1/2}\boldsymbol{\eta},$$

where The  $I(\cdot)$ 's are the corresponding Fisher information matrices.

- In large samples,  $\text{sd}(\theta|\mathbf{y}) \approx \text{diag}\{I_{\Theta}(\hat{\theta})^{-1/2}\}$  via the Bayesian CLT, so that  $I_{\Theta}(\theta)^{1/2}\theta$  may be viewed as a **scaled** or **standardized** version of  $\theta$ .

## The scale-transformed power prior (straPP)

- In most cases, we won't be able to solve for the transformation, and we will have to rely on non-linear equation solvers.
- An exception occurs when at least one of the parameter is from a normal model, e.g., if  $\eta$  is from a normal model,

$$I_H(\eta) = I_\eta = \mathbf{X}'_0 \mathbf{X}_0 / \sigma^2$$

and so

$$\eta = \sigma [\mathbf{X}'_0 \mathbf{X}_0]^{-1/2} [I_\Theta(\theta)]^{1/2} \theta$$

- If  $\eta$  (i.e., the historical parameter) is from a normal model, then there is a Jacobian adjustment:

$$\frac{\partial \eta}{\partial \theta'} = \sigma [\mathbf{X}'_0 \mathbf{X}_0]^{-1/2} \left\{ \frac{\partial I_\Theta^{1/2}}{\partial \theta'} \theta + I_\Theta(\theta)^{1/2} \right\}.$$

- **Note:** In many cases, the straPP transformation is not 1:1, but it is 1:1 in a subset of the parameter space.

## The scale-transformed power prior (straPP)

- More generally, suppose  $\theta$  = current data set parameter and suppose we can solve  $\theta = g(\eta)$  for some function  $g$ .
- The straPP is given by

$$\pi_{\text{straPP}}(\theta|D_0, a_0) = \pi_{\text{PP}}(g^{-1}(\theta)|D_0, a_0) \left| \frac{\partial g^{-1}}{\partial \theta} \right|$$

- If it is easier to solve  $\eta = h(\theta)$ , we can simply do the transformation after sampling  $\eta$  (**complementary sampling**)

$$p(\eta|D, D_0, a_0) \propto L(h^{-1}(\eta)|D)L(\eta|D_0)^{a_0}\pi_0(\eta).$$

Note that there is no Jacobian adjustment here.

## ESTEEM Trial: straPP

- The likelihoods for **normal** and **logistic** regression models are given by

$$L(\beta, \sigma^2 | D_j) \propto \prod_{z=0}^1 \prod_{i=1}^{n_{jz}} \sigma \exp \left\{ -\frac{1}{2\sigma^2} (y_{jzi} - \mathbf{x}'_i \beta)^2 \right\}$$

and

$$L(\gamma | D_j) \propto \prod_{i=1}^{n_j} \frac{\exp\{y_{ji} \mathbf{x}'_{ji} \gamma\}}{1 + \exp\{\mathbf{x}'_{ji} \gamma\}}$$

- This yields

$$I_B(\beta, \sigma^2 | D_j) = \frac{\mathbf{X}'_j \mathbf{X}_j}{\sigma^2}$$

and

$$I_C(\gamma | D_j) = \mathbf{X}'_j W(\gamma | X_j) \mathbf{X}_j,$$

where  $W(\gamma | X_j) = \text{diag}\{p_{ij}(1 - p_{ij}), i = 1, \dots, n_j\}$ ,  $p_{ij} = \text{logit}^{-1}(\mathbf{x}'_i \gamma)$

- Using the historical data  $D_0$ , we may solve for  $\beta$  via

$$\beta = \sigma(\mathbf{X}'_0 \mathbf{X}_0)^{-1/2} [\mathbf{X}'_0 W(\gamma | \mathbf{X}_0) \mathbf{X}_0]^{1/2} \gamma$$

- We use the **spectral decomposition** for the matrix square root, i.e.,

$$\mathbf{A}^{1/2} = \mathbf{P} \boldsymbol{\Lambda}^{1/2} \mathbf{P}',$$

where  $\boldsymbol{\Lambda} = \text{diag}\{\lambda_1, \dots, \lambda_p\}$  are eigenvalues and  $\mathbf{P}$  are the corresponding eigenvectors.

- We use **partial borrowing** techniques to avoid borrowing from the intercept.

## ESTEEM Trial: straPP

- When the straPP assumption is not tenable, one may use the generalized straPP (**gen-straPP**).
- The gen-straPP assumes

$$I_B(\beta)^{1/2}\beta = I_C(\gamma)^{1/2}\gamma + \mathbf{c}_0, \quad \mathbf{c}_0 \sim N_p(\mathbf{0}, \tau^2 I_p).$$

- For the ESTEEM trials, we may solve

$$\beta = \sigma^2 [\mathbf{X}'_0 \mathbf{X}_0]^{-1/2} \left\{ [\mathbf{X}'_0 \mathbf{W}(\gamma) \mathbf{X}_0]^{1/2} \gamma + \mathbf{c}_0 \right\}.$$

- We may set  $\tau$  hierarchically, e.g.,  $\tau \sim N^+(0, 1)$ . Alternatively, we may elicit  $\tau$ .

## ESTEEM Trial: straPP

- We implement the straPP and gen-straPP for the ESTEEM studies.
- R and Stan code for the implementation of the strapp and gen-straPP is available at  
[Examples/ESTEEM\\_straPP.Rmd](#).
- Click here for the data analysis results.

## Propensity score approaches

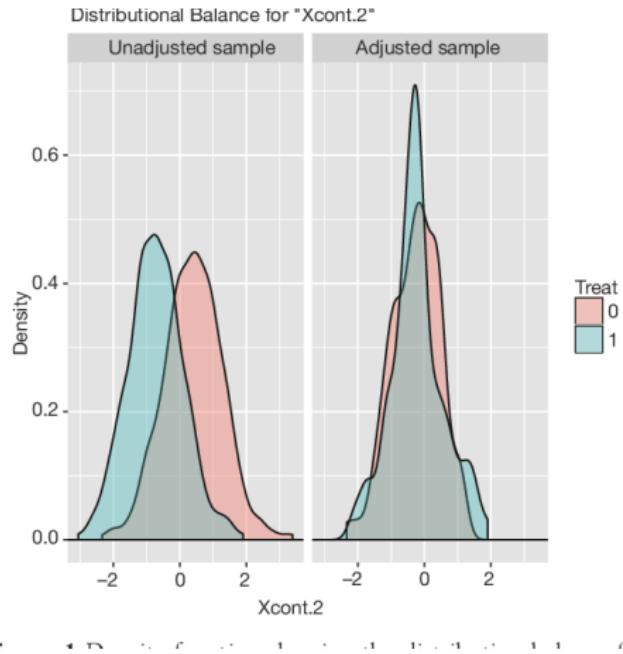
- Recently, so-called **propensity score integrated** priors have been proposed in the context of external controls.
- In the causal inference literature,  
 $\text{Propensity score}_i = \Pr(A_i = 1 | X = x)$ .
- The propensity score (PS) is a **balancing score**, i.e., subjects with similar propensity scores have similar characteristics (dimension reduction).
- In the context of external controls, the PS is

$$e_{ij} = \Pr(S_{ij} = 1 | X = x_{ij}), \quad j = 0, 1, \quad i = 1, \dots, n_j$$

where  $S_{ij} = j$  is the study ID of participant  $i$  in data set  $D_j$ .

- Propensity scores are about **study design**, not analysis.

# Propensity score as a balancing score



Distribution of covariate before and after PS adjustment.

## Propensity score approaches

- In the following three slides, we discuss three popular PS approaches as defined in the **causal inference** literature, i.e.,

$$\text{PS} = \Pr(\text{receive treatment} | X = x).$$

- For dealing with external controls, we may replace **receive treatment** with **belongs to current study**.
- It is important to note that the causal DAGs do not directly translate over to enrolling to the external controls literature.
  - e.g., Observational data DAG:  $X \rightarrow A \rightarrow Y$
  - Clinical trials:  $X \rightarrow S \rightarrow Y$  and  $A \rightarrow Y$  (no connection between  $X$  and  $A$ , no real connection between  $S$  and  $A$ ).

## Propensity score stratification (PSS)

- PSS involves the following steps:
  - ① Estimate the PS.
  - ② Stratify individuals into one of  $M$  strata based on the propensity score (e.g.,  $M = 5$  = quintiles).
  - ③ Use an intercept/treatment effect model within each stratum and a meta-analysis for the overall effect

$$g [E(y_i | a_i, m_i = m)] = \beta_{m0} + \beta_{m1} a_i, \quad m = 1, \dots, M$$

- The overall intercept and treatment effect may be estimated via a hierarchical prior e.g.,

$$\beta_{mk} \sim N(\beta_k, \tau_k^{-1}), \quad s = 1, \dots, S, k = 0, 1,$$

$$\beta_k \sim N(\beta_{0k}, \sigma_{0k}^2), \quad k = 0, 1,$$

$$\tau_k \sim \text{Gamma}(\alpha_{0k}, \text{rate} = \gamma_{0k}), \quad k = 0, 1.$$

## Propensity score matching (PSM)

- PSM involves the following steps:
  - ① Estimate the PS (e.g., through logistic regression).
  - ② Match cases historical data to current data based on similarities in the PS (e.g., 1-1 nearest neighbor matching w/ a caliper of 0.01).
    - ★ Unmatched subjects get discarded.
    - ★ This results in a new data set  $D^*$  of size  $n^* \leq n$ .
  - ③ Use an intercept/treatment effect model using  $D^*$

$$g [E(y_i | a_i)] = \beta_0 + \beta_1 a_i.$$

- It might be defensible to include all current data participants.

## Inverse-probability weighting (IPW)

- IPW uses the PS to balance baseline patient characteristics in the treated and untreated groups by weighting each individual by the inverse probability of receiving his/her actual treatment:

$$w_i = \frac{a_i}{\Pr(a_i = 1|\mathbf{x})} + \frac{1 - a_i}{\Pr(a_i = 0|\mathbf{x})} = \frac{a_i}{e_i} + \frac{1 - a_i}{1 - e_i}.$$

- The participant's likelihood contribution is then

$$L_i = [L(\theta|y_i)]^{w_i}.$$

## Criticisms of PS methods

- Cannot be fully Bayesian.

$$f(y, a, \mathbf{x} | \boldsymbol{\theta}) = f(y|a, \mathbf{x}; \boldsymbol{\theta}_Y) f(a|\mathbf{x}; \boldsymbol{\theta}_A) f(\mathbf{x} | \boldsymbol{\theta}_X).$$

Thus if  $\pi(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta}_Y)\pi(\boldsymbol{\theta}_A)\pi(\boldsymbol{\theta}_X)$ ,

$$p(\boldsymbol{\theta}|D) = p(\boldsymbol{\theta}_Y|D)p(\boldsymbol{\theta}_A|D)p(\boldsymbol{\theta}_X|D)$$

- Uncertainties ignored in PS (can do quasi-Bayes to remedy this):
  - ① Draw PS; then stratify.
  - ② Draw  $\boldsymbol{\theta}_Y$  based on current.
- Can be difficult to define the study population at the end of the study.
- In the context of external controls, underlying assumption is

$$f(\mathbf{x}) = f(\mathbf{x}_0) \implies f(y|\mathbf{x}) = f(y_0|\mathbf{x}_0),$$

but we really only care about the r.h.s.

# Propensity score integrated power prior (PSIPP)

- The PSIPP [15, 16] consists of two steps.

## ① Design stage:

- ★ Estimate propensity score  $e_{ij} = \Pr(S_{ij} = 1 | x_{ij})$ ,  $j = 0, 1$ ,  $i = 1, \dots, n_j$
- ★ **Trim:** Remove individuals from historical data set whose PS is not in the range of the current, i.e., keep only observations in  $[e_{(1)1}, e_{(n_1)1}]$
- ★ Place remaining observations into  $M$  strata using  $M$ -tiles of the PS, resulting in  $D_0 = (D_{01}, \dots, D_{0M})$ .

## ② Analysis stage:

- ★ Analyze the data using the prior
$$\pi_{\text{PSIPP}}(\theta | D_0, a_0) = \prod_{m=1}^M L(\theta_m | D_{0m})^{a_{0m}} \pi_0(\theta_m).$$
- ★ Authors recommend taking a weighted average of the treatment effects.

# Propensity score integrated power prior (PSIPP)

- There are two key **tuning parameters**:
  - ① How many strata ( $M$ )?
  - ② How to choose  $a_0 = (a_{01}, \dots, a_{0M})$ ?
- The authors recommend  $a_{0m} = \frac{\lambda_{00m}}{n_{00m}}$ , where
  - ▶  $\lambda_{00m} = \min \left\{ \frac{\tilde{n}_{00} r_m}{\sum_{l=1}^M r_l}, n_{00m} \right\}$  = number of historical controls to borrow from the  $m^{th}$  stratum.
  - ▶  $\tilde{n}_{00} \leq \min\{n_{00}, n_{11} - n_{10}\}$  = total number of historical controls to borrow.
  - ▶  $n_{00m}$  = number of historical controls assigned to  $m^{th}$  stratum.
  - ▶  $r_m$  = measure of similarity between patients in  $m^{th}$  stratum, e.g., overlap in PS within  $m^{th}$  stratum.

## ESTEEM Trial: PSIPP

- We apply the PSIPP to the simulated ESTEEM trials data sets.
- R and Stan code for the implementation of the PSIPP is available at [Examples/ESTEEM\\_psipp.Rmd](#)
- [Click here for the data analysis results.](#)

# The Latent Exchangeability Prior (LEAP)

- Thus far, we have discussed “**static**” borrowing priors and “**dynamic**” borrowing priors.
  - ▶ **Static**—amount borrowing is fixed (e.g., power prior)
  - ▶ **Dynamic**—amount of borrowing is data-driven (e.g., normalized power prior, BHM)
- Typically, **static** priors are easier to implement, but have poor type I error rates.
- Conversely, **dynamic** priors typically have better type I error rates, but can be more difficult to implement.
- The LEAP is a **dynamic** borrowing prior that can be easily implemented in any MCMC software.

# The Latent Exchangeability Prior (LEAP)

- The LEAP (Alt et al, forthcoming) is predicated on two assumptions:
  - ➊ The historical data are generated from a mixture model,

$$f(y_0|\boldsymbol{\theta}, \gamma) = \sum_{k=1}^K \gamma_k f_k(y_0|\boldsymbol{\theta}_k) = \gamma_1 f_1(y_0|\boldsymbol{\theta}_1) + \sum_{k=2}^K \gamma_k f_k(y_0|\boldsymbol{\theta}_k). \quad (7)$$

where  $\gamma_k \in (0, 1)$  and  $\sum_{k=1}^K \gamma_k = 1$

- ➋ The current data are generated from one of these components (we assume WLOG that it is the first one)

$$f(y|\boldsymbol{\theta}) = f_1(y|\boldsymbol{\theta}_1).$$

- Given  $n_0$  observations of historical data with proper initial prior  $\pi_0(\boldsymbol{\theta})$ , the LEAP is the prior induced by the mixture model in (7).

# The Latent Exchangeability Prior (LEAP)

- Note that we may alternatively express the mixture model in (7) in **latent class** form as

$$f(y_{0i}, c_{0i} | \boldsymbol{\theta}, \gamma) = \prod_{k=1}^K \{\gamma_k f_k(y_{0i} | \boldsymbol{\theta}_k)\}^{1\{c_{0i}=k\}}, \quad (8)$$

where  $c_{0i} \in \{1, \dots, K\}$  indexes the latent class to which subject  $i$  belongs.

- Thus, historical data participants with  $c_{0i} = 1$  have special meaning. Namely, they are **exchangeable** (i.e., have the same parameters) as the current data participants.
- Moreover,  $\gamma_1 = \Pr(c_{0i} = 1) = \Pr(\text{exchangeable with current data}).$

## LEAP: Computational Details

- For any MCMC suite that allows for sampling of discrete parameters (e.g., Nimble, JAGS, SAS PROC MCMC), it is straightforward to implement the LEAP via the **latent class representation** (8).
- For Stan, we must use the marginalized representation in (7). We can generate the latent classes ex-post via

$$p(\boldsymbol{\theta}, \boldsymbol{\gamma}, \mathbf{c} | D, D_0) = p(\mathbf{c} | \boldsymbol{\gamma}, \boldsymbol{\theta}; D, D_0)p(\boldsymbol{\theta}, \boldsymbol{\gamma} | D, D_0).$$

- There is a single tuning parameter  $K \geq 2$ . Typically,  $K = 2$  suffices, but the following may be helpful at selecting  $K$ :
  - ① Deviance Information Criterion (DIC) (Spiegelhalter et al, 2002) [17]
  - ② Bayes factor (BF)
  - ③ Leave-One-Out Information Criterion (LOO-IC; Vehtari et al, 2017 [18])

## ESTEEM: LEAP Example

- We implement the LEAP with  $K = 2$ .
- We borrow only from the controls, for which we have  $n_{00} = 282$  and  $n_{01} = 137$  patients assigned to placebo in the ESTEEM-I and ESTEEM-II trials, respectively.
- We elicit
  - ▶  $\beta_j \sim N(0, 10^2)$
  - ▶  $\sigma \sim \text{Half-Cauchy}(0, 1)$
  - ▶  $\gamma \sim U(0, 0.49 \approx n_{01}/n_{00})$

## ESTEEM: LEAP Example

- We apply the LEAP to the simulated ESTEEM trials data sets.
- R and Stan code for the implementation of the LEAP is available at [Examples/ESTEEM\\_leap.Rmd](#)
- [Click here for the data analysis results.](#)

Informative prior elicitation from other sources

## Informative prior elicitation from other sources

- Sometimes, we may wish to use one of the informative priors, but we do not have access to the underlying data.
- In this section, we will discuss informative prior elicitation based on:
  - ① Expert opinion.
  - ② Summary statistics.

## Conjugate priors for GLMs

- Recall that i.i.d. exponential family models admit conjugate priors of the form

$$\pi_{\text{DY}}(\theta | \lambda_0, \mu_0) = \frac{1}{C(\lambda, \mu_0)} \exp \{ \lambda_0 [\mu_0 \theta - b(\theta)] \},$$

where

- $\lambda_0 \in \{0, 1, \dots, n\}$  = precision parameter = prior sample size.
- $\mu_0$  = location parameter = prediction for  $E(y)$ .

## Conjugate priors for GLMs

- Chen and Ibrahim (2003) [19] extended the conjugate prior to GLMs

$$\pi_{\text{CI}}(\boldsymbol{\beta} | \lambda_0, \boldsymbol{\mu}_0) = \frac{1}{C(\lambda_0, \boldsymbol{\mu}_0)} \exp \left\{ \lambda_0 [\theta(\mathbf{X}\boldsymbol{\beta})' \boldsymbol{\mu}_0 - \mathbf{1}' b(\theta(\mathbf{X}\boldsymbol{\beta}))] \right\},$$

where

- $\lambda_0 \in [0, 1]$  = discounting factor.
- $\boldsymbol{\mu}_0$  = location parameter = prediction for  $E(\mathbf{y}|\mathbf{X})$ .
- $\mathbf{X}$  = design matrix for **current data**.
- $\theta(\cdot) = (\dot{b}^{-1} \circ g^{-1})(\cdot)$  =  $\theta$ -link function, where  $g(\cdot) = \mu$ -link function.  
For canonical link functions,  $\theta(\eta) = \eta$ .
- $\mathbf{1} = (1, \dots, 1)'$

## Conjugate priors for GLMs

- Note if  $\mathbf{X} = \mathbf{X}_0$ ,  $\lambda_0 = a_0$ , and  $\mu_0 = \mathbf{y}_0$ , the CI prior is a power prior.
- For normal linear models, the CI prior is  $N\left([\mathbf{X}'\mathbf{X}]^{-1}\mathbf{X}'\boldsymbol{\mu}_0, \frac{[\mathbf{X}'\mathbf{X}]^{-1}}{\lambda_0}\right)$ .
- The hyperparameter  $\lambda_0$  may be chosen the same way as  $a_0$  for the power prior. The ESS is given by  $ESS = \lambda_0 \times n$ .
- The hyperparameter  $\boldsymbol{\mu}_0$  may be chosen as expert opinion (e.g., prediction for the mean response) or from summary statistics of a previous study (e.g.,  $\mu_{0i} = g^{-1}(\mathbf{x}_i'\hat{\boldsymbol{\beta}}_0)$ ).

## Conjugate priors for GLMs: ESTEEM Studies

- We implement the conjugate prior using the outcome:

$y_i$  = reduction of at least 75% from baseline in PASI score.

- We pretend that we do not have the full historical data set, but rather only have the MLE and standard errors.
- We elicit  $\mu_0 = \text{logit}^{-1}(\mathbf{X}\hat{\beta}_0)$  and  $\lambda_0 = 0.5 \times \min\left\{1, \frac{n}{n_0}\right\}$ .
- For simplicity, we will borrow from historical controls **and** treatment.

# Conjugate priors for GLMs: ESTEEM Studies

- R and Stan code for the implementation of the PSIPP is available at [Examples/ESTEEM\\_conjugatePriorGLMs.Rmd](#)
- Click here for the data analysis results.

## The Hyper- $g$ Prior

- Similar to the normalized power prior, we can avoid elicitation of  $\lambda_0$  by taking it as random, yielding the Hyper- $g$  prior.
- Of course, this comes at the cost of having to estimate the normalizing constant.
- Note that the prior mode of  $\pi_{\text{CI}}$  is  $\hat{\beta}_{\mu_0} = \text{MLE}$  treating  $\mu_0$  as data.
- We may thus use a **Laplace approximation** to estimate the normalizing constant.
- Note that since the CI prior is proportional to a GLM, we may use IRLS to estimate  $\hat{\beta}_{\mu_0}$ , and the Hessian is equal to the observed Fisher information matrix.

# The Hyper- $g$ Prior: Laplace Approximation

- Let  $\pi_{\text{CI}}^*(\beta|\lambda_0, \mu_0)$  denote the unnormalized prior density. We can compute a second order Taylor expansion about  $\hat{\beta}_{\mu_0}$

$$\log \pi_{\text{CI}}^*(\beta|\lambda, \mu_0) \approx \log \pi_{\text{CI}}^*(\hat{\beta}_{\mu_0}|\lambda, \mu_0) - \frac{1}{2}(\beta - \hat{\beta}_{\mu_0})' \mathcal{J}(\beta - \hat{\beta}_{\mu_0}),$$

where

$$\mathcal{J} = \mathcal{J}(\hat{\beta}_{\mu_0}|\lambda, \mu_0) = -\frac{\partial^2 \log \pi_{\text{CI}}}{\partial \beta \partial \beta'} = \lambda \times (\text{obs. Fisher info.})$$

- It follows that

$$Z(\lambda, \mu_0) = \int \pi_{\text{CI}}^*(\beta|\lambda, \mu_0) d\beta \approx \pi_{\text{CI}}^*(\hat{\beta}_{\mu_0}|\lambda, \mu_0) \int \exp \left\{ -\frac{1}{2}(\beta - \hat{\beta}_{\mu_0})' \mathcal{J}(\beta - \hat{\beta}_{\mu_0}) \right\} d\beta$$

- Note  $\int \exp\{\text{integrand}\} d\beta$  is the normalizing constant of a  $N_p(\hat{\beta}_{\mu_0}, \mathcal{J}^{-1})$  density, so the approximation is given by

$$Z(\lambda, \mu_0) \approx \pi_{\text{CI}}^*(\hat{\beta}_{\mu_0}|\lambda, \mu_0) (2\pi)^{p/2} |\mathcal{J}|^{-1/2}$$

## The Hierarchical Prediction Prior (HPP)

- The prior prediction  $\mu_0$  in the CI conjugate prior is typically elicited on the basis of **expert opinion** or **summary statistics**.
- There is uncertainty surrounding the accuracy of this prediction. One way to account for this uncertainty is through the precision parameter  $\lambda_0$ .
- However,  $\lambda_0$  cannot simultaneously control for the level of informativeness in the prior **and** the uncertainty surrounding the prior prediction.

## The Hierarchical Prediction Prior (HPP)

- For example, in an i.i.d. Bernoulli( $\mu$ ) model, the CI prior is the DY prior with  $n_0 = \lambda n$  and  $\bar{y}_0 = n^{-1} \sum_{i=1}^n \mu_{0i}$ , i.e.,

$$\pi_{\text{DY}}(\mu | n_0, \bar{y}_0) = \frac{1}{B(n_0 \bar{y}_0, n_0(1 - \bar{y}_0))} \mu^{n_0 \bar{y}_0 - 1} (1 - \mu)^{n_0(1 - \bar{y}_0) - 1},$$

which is a Beta( $n_0 \bar{y}_0, n_0(1 - \bar{y}_0)$ ) prior.

- Suppose an expert provides  $\bar{y}_0 = 0.3$  and says they are 95% confident the true value of  $\mu$  is between  $(0.2, 0.4)$ .
- To find the appropriate  $n_0$ , we must solve

$$\Pr(0.2 < \mu < 0.4 | n_0, \bar{y}_0 = 0.3) = 0.95,$$

yielding  $n_0 = 78.8$ .

## The Hierarchical Prediction Prior (HPP)

- This constrains the level of informativeness of the prior. Namely, if we choose any  $n_0 \neq 78.8$ , we will no longer have our uncertainty about the prediction implemented.
- Thus, we see that the DY and CI priors result in a tradeoff between
  - ▶ Controlling for how impactful the prior will be in the resulting posterior density.
  - ▶ Incorporating uncertainty surrounding the prior prediction.

# The Hierarchical Prediction Prior (HPP)

- The hierarchical prediction prior (HPP) of Alt et al. (2022) [20] introduces a hierarchical component for the prior prediction, i.e.,

$$\pi_{\text{HPP}}(\mu|n_0, \bar{y}_0, \tau_0) = \pi_{\text{DY}}(\mu|n_0, m)\pi_{\text{HPP}}(m|\tau_0, \bar{y}_0)dm,$$

where  $\pi_{\text{HPP}}(m|\tau_0, \bar{y}_0)$  is a **hyperprior** over the range  $E(y)$  with mean  $\bar{y}_0$ .

- For a Bernoulli model, we may specify  $m \sim \text{Beta}(\alpha_0, \beta_0)$  where  $\frac{\alpha_0}{\alpha_0 + \beta_0} = \bar{y}_0$ , yielding

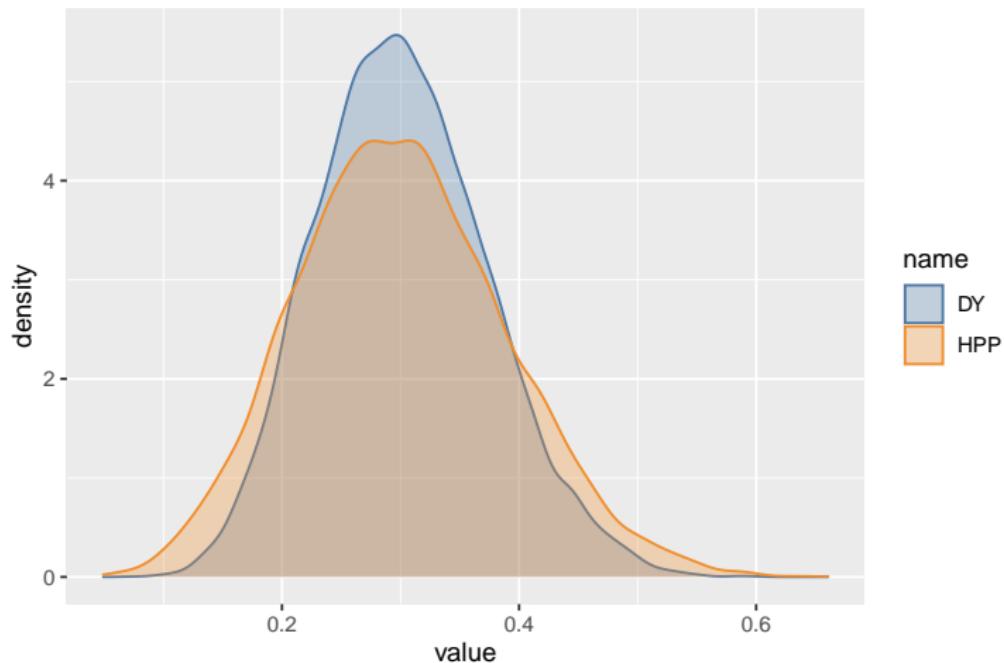
$$\pi_{\text{HPP}}(\mu|n_0, \bar{y}_0, \tau_0) = \int \frac{\mu^{n_0 m - 1} (1 - \mu)^{n_0(1-m)-1}}{B(n_0 m, n_0(1-m))} \frac{m^{\alpha_0-1} (1-m)^{\beta_0-1}}{B(\alpha_0, \beta_0)} dm,$$

- For exponential family models, it can be shown that, still,  $\mu_0 = E(y)$ .

## The Hierarchical Prediction Prior (HPP)

- Suppose now we wish to design a study where an expert believes with 95% probability that  $\mu \in (0.2, 0.4)$ .
- The planned study size is  $n = 40$ . Thus, it would be inappropriate to use the DY prior.
- Conversely, we may use the HPP with
  - ▶  $m \sim \text{Beta}(\alpha_0 = 78.8 \times 0.3, \beta_0 = 78.8 \times (1 - 0.3))$
  - ▶  $n_0 = 40$

# The Hierarchical Prediction Prior (HPP)



Comparison of HPP and DY priors when  $n_0 = 40$ ,  $\alpha_0 = 78.8 \times 0.3$ ,  $\beta_0 = 78.8 \times (1 - 0.3)$ .

# The Hierarchical Prediction Prior (HPP)

- For regression models, the HPP is given by

$$\pi_{\text{HPP}}(\boldsymbol{\beta}|\lambda, \boldsymbol{\tau}_0, \boldsymbol{\mu}_0) = \int \pi_{\text{CI}}(\boldsymbol{\beta}|\lambda, \mathbf{m}) \pi_{\text{HPP}}(\mathbf{m}|\boldsymbol{\tau}_0, \boldsymbol{\mu}_0),$$

where  $\mu_{0i} = E(y_i|\mathbf{X} = \mathbf{x}_i)$  and

$$\pi_{\text{HPP}}(\mathbf{m}|\boldsymbol{\tau}_0, \boldsymbol{\mu}_0) \propto \prod_{i=1}^n \exp \left\{ \tau_{0i} \left[ b^{-1} \left( m_i \mu_{0i} - b(b^{-1}(m_i)) \right) \right] \right\}$$

- Example:** Poisson regression model:

$$\pi_{\text{HPP}}(\mathbf{m}|\lambda_0, \boldsymbol{\mu}_0) \propto \prod_{i=1}^n m_i^{\tau_{0i}\mu_{0i}-1} \exp\{-\tau_{0i}m_i\}$$

## The Hierarchical Prediction Prior (HPP)

- The posterior distribution of  $\mathbf{m}$  plays an important role, namely,

$$p(\boldsymbol{\beta}|D) = \int p(\boldsymbol{\beta}|D, \mathbf{m}; \lambda) p(\mathbf{m}|D; \mu_0, \tau_0) d\mathbf{m}$$

- Thus, the posterior distribution of the HPP is that using the CI prior **averaged over** the posterior distribution of the predictions, which is itself modified by the data.
- This adds an element of robustness—namely, discounting is conducted **at the individual level**.
- For a normal linear model, one can show that

$$E(m_i|D) = \lambda_i \mathbf{x}'_i \hat{\boldsymbol{\beta}} + (1 - \lambda_i) \mu_{0i}, \quad \lambda_i \in (0, 1)$$

where  $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y}$  is the least squares estimator of  $\boldsymbol{\beta}$ .

# ESTEEM Studies: The Hierarchical Prediction Prior (HPP)

- We implement the HPP for the ESTEEM logistic regression model.
- We assume that we only know **summary statistics** of the historical data.
  - ① The MLE  $\hat{\beta}_0$ .
  - ② The covariance matrix  $\hat{\Sigma}_0$ .
- The **hierarchical prior** is a beta prior with:
  - ▶ Mean:  $\mu_{0i} = \frac{e^{x_i' \hat{\beta}_0}}{1+e^{x_i' \hat{\beta}_0}}$
  - ▶ Variance:  $\sigma_{0i}^2 = [\mu_{0i}(1 - \mu_{0i})]^2 x_i' \hat{\Sigma}_0 x_i$  (via the **delta method**).

# ESTEEM Studies: The Hierarchical Prediction Prior (HPP)

- Note that we must estimate the normalizing constant of the CI prior, which we do via the **Laplace approximation** previously derived:

$$Z(\lambda, \mathbf{m}) \approx \pi_{\text{CI}}^* \left( \hat{\beta}_{\mathbf{m}} | \lambda, \boldsymbol{\mu}_0 \right) (2\pi)^{p/2} |\mathcal{J}(\lambda, \mathbf{m})|^{-1/2},$$

where

- ▶  $\pi_{\text{CI}}^*$  = unnormalized prior density
- ▶  $\hat{\beta}_{\mathbf{m}}$  = MLE treating  $\mathbf{m}$  as outcome.
- ▶  $\mathcal{J}(\lambda, \mathbf{m}) = \lambda \times \mathbf{X}' \mathbf{W} \mathbf{X}$ ,  $\mathbf{W} = \text{diag}\{m_i(1 - m_i), i = 1, \dots, n\}$

# ESTEEM Studies: The Hierarchical Prediction Prior (HPP)

- R and Stan code for the implementation of the HPP is available at [Examples/ESTEEM\\_HPP.Rmd](#)
- Click here for the data analysis results.

## Part 3: Case Studies in Prior Elicitation

## Case Study 1: Power Prior - AIDS Data

## Case Study 1: Power Prior - AIDS Data

- The ACTG019 study was a double blind placebo-controlled clinical trial comparing zidovudine (AZT) to placebo in persons with CD4 counts less than 500.
- The sample size for this study, excluding cases with missing data, was  $n_0 = 823$ .
- The response variable ( $y_0$ ) for these data is binary with a 1 indicating death, development of AIDS or development of AIDS-related complex (ARC), and a 0 indicates otherwise.
- The ACTG036 study was also a placebo-controlled clinical trial comparing AZT to placebo in patients with hereditary coagulation disorders.
- The sample size in this study (excluding missing data) was  $n = 183$ .
- The response variable ( $y$ ) for these data is binary with a 1 indicating death, development of AIDS or ARC, and 0 otherwise.

# Case Study 1: Power Prior - AIDS Data

Table 1. Trial Experience of All Subjects Evaluated.

TREATMENT STATUS	STUDY GROUP			ALL
	PLACEBO	500-MG ZIDOVUDINE	1500-MG ZIDOVUDINE	
No. of subjects	428	453	457	1338
Mean weeks of follow-up*	61	55	51	55
<i>no. of subjects</i>				
Ineligible in minor respects	16	15	12	43
Eligible but never started treatment	7	4	6	17
Withdrew voluntarily from drug	111	73	81	265
Because of medical symptoms	8	13	23	44
Because of inconvenience†	28	30	36	94
To seek other therapy	50	15	9	74
No stated reason	25	15	13	53
Lost to follow-up after starting treatment	41	19	27	87

\*Measured from date of randomization to August 10, 1989.

†Includes subjects unwilling or unable to continue scheduled clinical visits.

# Case Study 1: Power Prior - AIDS Data

Table 5. Changes from Base-Line HIV p24 Antigen Levels.\*

TIME/ANTIGEN MEASURE	STUDY GROUP		
	PLACEBO	500-MG ZIDOVUDINE	1500-MG ZIDOVUDINE
<b>Week 8</b>			
No. with decrease/no. of samples (%)	2/10 (20)	7/14 (50)	4/10 (40)
Antigen level (pg/ml)			
Median	51	12	44
Median change	+1	-21	-44
<b>Week 16</b>			
No. with decrease/no. of samples (%)	5/24 (21)	19/35 (54)	20/27 (74)
Antigen level (pg/ml)			
Median	45	35	18
Median change	+2	-20	-29
<b>Week 32</b>			
No. with decrease/no. of samples (%)	2/14 (14)	11/25 (44)	7/17 (41)
Antigen level (pg/ml)			
Median	42	40	25
Median change	+14	-24	-18
<b>Week 48</b>			
No. with decrease/no. of samples (%)	3/13 (23)	8/22 (36)	6/15 (40)
Antigen level (pg/ml)			
Median	64	38	37
Median change	+12	-10	-11
Odds ratio†	—	3.6	4.9
P value‡	—	0.017	0.013

\*For each set of measurements, the number of subjects in whom the HIV p24 antigen level decreased by  $\geq 50$  percent of the base-line value at the indicated time is shown, followed by the number of subjects who provided a specimen available for assay at that time in whom HIV p24 antigen was detectable ( $> 10$  pg per milliliter) at base line. The number in parentheses is the percentage of subjects who had a  $\geq 50$  percent decrease.

†For the comparison of the placebo group to the respective zidovudine groups.

‡Derived according to the method of Wei and Johnson.<sup>22</sup>

## Results: ACTG 019 Study

## Case Study 1: Power Prior - AIDS Data

- Recall that the power prior is given by  $\pi(\beta) \propto \mathcal{L}(\beta|D_0)^{a_0} \pi_0(\beta)$ .
- We take  $\pi_0(\beta) \propto 1$ .
- We include CD4 count, age, and treatment as covariates.
- We consider  $a_0 \in \{0.00, 0.415, 1.00\}$ .
- The value  $a_0 = 0.415$  was chosen to mimic a Bayesian hierarchical model.

## Case Study 1: Power Prior - AIDS Data

$a_0$	Parameter	Posterior Mean	Posterior Std Dev	95% HPD Interval
0	$\beta_0$	-4.781	0.849	(-6.461, -3.223)
	$\beta_1$	-1.636	0.449	(-2.539, -0.791)
	$\beta_2$	0.122	0.234	(-0.334, 0.587)
	$\beta_3$	-0.057	0.380	(-0.802, 0.698)
0.415	$\beta_0$	-3.196	0.253	(-3.691, -2.708)
	$\beta_1$	-0.779	0.175	(-1.121, -0.434)
	$\beta_2$	0.259	0.142	(-0.026, 0.531)
	$\beta_3$	-0.344	0.196	(-0.724, 0.043)
1	$\beta_0$	-3.041	0.169	(-3.379, -2.722)
	$\beta_1$	-0.677	0.123	(-0.917, -0.437)
	$\beta_2$	0.302	0.110	( 0.083, 0.513)
	$\beta_3$	-0.377	0.139	(-0.654, -0.109)

Posterior summary of the power prior with the AIDS data sets.  $\beta_1$ ,  $\beta_2$ , and  $\beta_3$  respectively refer to CD4 count, age, and treatment

## Case Study 1: Power Prior - AIDS Data

- We compare the results of the power prior with the Bayesian hierarchical model (BHM) (i.e., the MAP prior).
- We elicit
  - ▶  $\beta, \beta_0 \sim N_4(\mu, \Omega)$ ,
  - ▶  $\pi(\mu) \propto 1$ ,
  - ▶  $\Omega = \text{diag}\{\omega_0, \dots, \omega_3\}, \quad \omega_j \sim \text{IG}(1, 0.005)$ .

## Case Study 1: Power Prior - AIDS Data

Parameter	Posterior Mean	Posterior Std Dev	Parameter	Posterior Mean	Posterior Std Dev
$\beta_0$	-3.128	0.238	$\beta_{00}$	-3.044	0.177
$\beta_1$	-0.728	0.161	$\beta_{01}$	-0.671	0.129
$\beta_2$	0.261	0.138	$\beta_{02}$	0.323	0.118
$\beta_3$	-0.336	0.184	$\beta_{03}$	-0.387	0.144
$\mu_0$	-3.083	0.224	$\Omega_{00}$	0.031	0.227
$\mu_1$	-0.702	0.170	$\Omega_{11}$	0.021	0.091
$\mu_2$	0.293	0.149	$\Omega_{22}$	0.020	0.134
$\mu_3$	-0.361	0.179	$\Omega_{33}$	0.021	0.104

Posterior summary of the BHM with the AIDS data sets.  $\beta_1$ ,  $\beta_2$ , and  $\beta_3$  respectively refer to CD4 count, age, and treatment

## Case Study 2: Commensurate Prior - Colorectal Cancer Data

## Case Study 2: Commensurate Prior - Colorectal Cancer Data

- We illustrate the commensurate prior (CP) using data from two successive randomized controlled colorectal cancer clinical trials.
- The initial trial (Saltz et al. 2000) randomized  $n_0 = 683$  patients with previously untreated metastatic colorectal cancer between May 1996 and May 1998 to one of three regimens:
  - ① Irinotecan alone
  - ② Irinotecan and bolus Fluorouracil plus Leucovorin (IFL)
  - ③ Fluorouracil and Leucovorin (5FU/LV) ("standard therapy")
- IFL resulted in significantly longer progression-free survival and overall survival than both Irinotecan alone and 5FU/LV and became the standard of care treatment.

## Case Study 2: Commensurate Prior - Colorectal Cancer Data

- The subsequent trial (Goldberg et al. 2004) compared three new (at the time) drug combinations in  $n = 795$  patients with previously untreated metastatic colorectal cancer, randomized between May 1999 and April 2001.
- Patients in the first drug group received the current “standard therapy,” the IFL regimen identical to that used in the historical study.
- The second group received Oxaliplatin and infused Fluorouracil plus Leucovorin (abbreviated FOLFOX)
- The third group received Irinotecan and Oxaliplatin (abbreviated IROX); both of these latter two regimens were new as of the beginning of the second trial.

## Case Study 2: Commensurate Prior - Colorectal Cancer Data

- Both trials recorded two bi-dimensional measurements on each tumor for each patient at regular cycles.
- The Saltz trial measured patients every 6 weeks for the first 24 weeks and every 12 weeks thereafter until death or disease progression
- The Goldberg trial measured every 6 weeks for the first 42 weeks, or until death or disease progression.
- The authors computed the sum of the longest diameter in cm ("ld sum") for up to 9 tumors for each patient at baseline.
- In both trials, disease progression was defined as a 25% or greater increase in measurable tumor or the appearance of new lesions.

## Case Study 2: Commensurate Prior - Colorectal Cancer Data

- Both trials recorded two bi-dimensional measurements on each tumor for each patient at regular cycles.
- The Saltz trial measured patients every 6 weeks for the first 24 weeks and every 12 weeks thereafter until death or disease progression
- The Goldberg trial measured every 6 weeks for the first 42 weeks, or until death or disease progression.
- The authors computed the sum of the longest diameter in cm ("ld sum") for up to 9 tumors for each patient at baseline.
- In both trials, disease progression was defined as a 25% or greater increase in measurable tumor or the appearance of new lesions.

## Case Study 2: Commensurate Prior - Colorectal Cancer Data

- The historical data consists of the IFL treatment arm from the initial study
- The current data consists of patients randomized to IFL or FOLFOX in the subsequent trial.
- We omit data from the Irinotecan alone and 5FU/LV arms in the Saltz study, and the IROX arm in the Goldberg study. The model incorporates baseline  $\text{ld sum}$  as a predictor.
- A parametric Weibull regression model is specified to compare disease progression among the FOLFOX and IFL regimens.

## Case Study 2: Commensurate Prior - Colorectal Cancer Data

- A Weibull accelerated failure time (AFT) model is given by

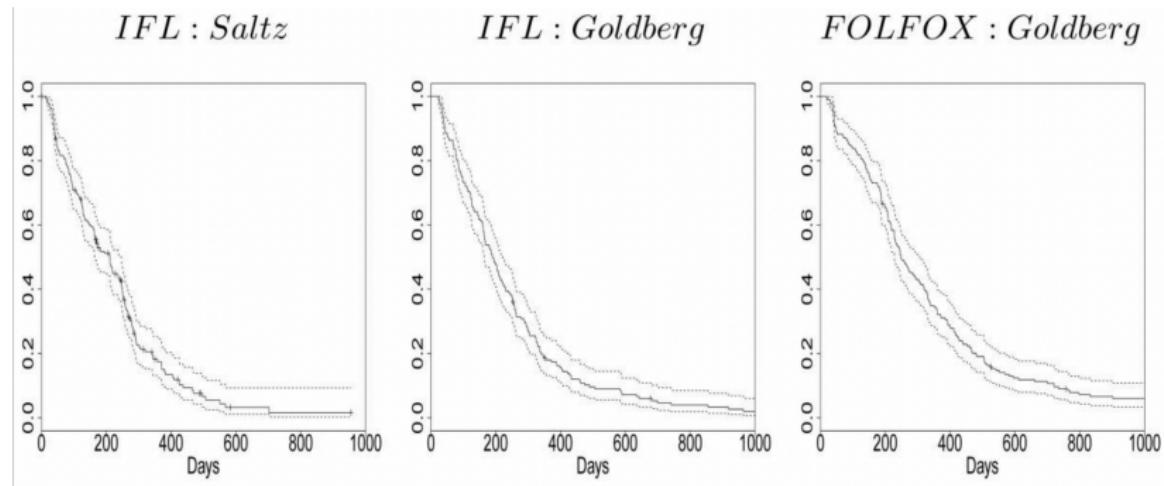
$$\tilde{y} = \log y = \mathbf{X}\boldsymbol{\beta} + \mathbf{z}\lambda + \sigma\epsilon,$$

where

- ①  $y$  = observed failure time
- ②  $\mathbf{X}$  = covariate matrix
- ③  $\mathbf{z}$  = vector of treatment indicators
- ④  $\lambda$  = treatment effect
- ⑤  $\sigma$  = AFT scale parameter
- ⑥  $\epsilon \sim \text{Gumbel}(0, 1)$  (log of Weibull), with PDF

$$f(\epsilon|\mu, \gamma) = \frac{1}{\beta} \exp \left\{ -\frac{\epsilon - \mu}{\gamma} - \exp \left( -\frac{\epsilon - \mu}{\gamma} \right) \right\}.$$

## Case Study 2: Commensurate Prior - Colorectal Cancer Data



Kaplan-Meier curves for time to disease progression. The plots suggest that the time to progression experience for subjects on IFL was similar in both the Saltz (left panel) and Goldberg trials (center), and that FOLFOX (right) is associated with somewhat prolonged time-to-progression.

## Case Study 2: Commensurate Prior - Colorectal Cancer Data

Separate analyses				
	Historical		Current	
	est	sd	est	sd
Intercept	5.503	0.058	5.555	0.067
BL ld sum	-0.043	0.051	-0.115	0.045
FOLFOX	-	-	0.417	0.092
$\log(\sigma)$	-0.291	0.060	-0.153	0.039

Results from running separate analyses of the current and historical data sets. FOLFOX is the experimental treatment, which was not an arm in the historical study.

## Case Study 2: Commensurate Prior - Colorectal Cancer Data

- Recall that the commensurate prior is given by

$$\pi(\boldsymbol{\beta}, \boldsymbol{\beta}_0, \sigma^2 | \boldsymbol{\tau}) \propto N(\boldsymbol{\beta} | \boldsymbol{\beta}_0, \text{diag}\{\boldsymbol{\tau}\}) \mathcal{L}(\boldsymbol{\beta}_0, \sigma^2 | D_0).$$

- For  $\boldsymbol{\tau}$ , the authors try three different approaches
  - Empirical Bayesian (EB) approach (point estimation).
  - Spike and slab prior (mentioned previously):
$$\pi(\tau) = 0.99 \times U(\tau | 0.005, 2) + 0.01 \times 1\{\tau = 200\}$$
  - $\tau_j \sim \text{Gamma}(1, 0.01)$

## Case Study 2: Commensurate Prior - Colorectal Cancer Data

	Full Model					
	EB		spike & slab		Gamma(1, 0.01)	
	est	sd	est	sd	est	sd
Intercept	5.541	0.054	5.547	0.058	5.546	0.058
BL ld sum	-0.100	0.040	-0.103	0.042	-0.105	0.042
FOLFOX	0.435	0.085	0.431	0.086	0.432	0.085
$\log(\sigma)$	-0.152	0.038	-0.158	0.038	-0.158	0.038
$\tau_1$	200	-	153.2	84.4	124.8	107.3
$\tau_2$	200	-	153.8	83.9	123.8	106.7
$\tau_3$	40.0	-	126.1	96.1	102.5	93.7

MCMC results of the commensurate prior applied to the colorectal cancer data sets.

## Case Study 2: Commensurate Prior - Colorectal Cancer Data

- Incorporating the Saltz data into the analysis of the Goldberg trial using the commensurate prior leads to more precise parameter estimates (i.e., reductions to the posterior standard deviation for the FOLFOX effect of nearly 9%, 8%, and 5% for the EB, spike and slab, and gamma models, respectively).
- These models provide considerably less borrowing of strength than that provided by pooling, which facilitates a 16% reduction in posterior standard deviation for estimating the FOLFOX effect.

## Case Study 3: Meta-analytic predictive Prior

## Case Study 3: Meta-analytic predictive Prior

- Recall that the meta-analytic predictive (MAP) prior is given by

$$\pi(\beta_1, \beta_0) \propto N(\beta_1 | \mu, \Sigma) \left\{ \prod_{j=1}^J \mathcal{L}(\beta_{0j} | D_{0j}) N(\beta_{0j} | \mu, \Sigma) \right\} \pi(\mu, \Sigma),$$

where

- ①  $\beta_1$  = current data parameter
- ②  $\beta_{0j}$  = parameter for  $j^{th}$  historical study
- ③  $\mu$  = meta-analytic mean
- ④  $\Sigma$  = covariance matrix governing heterogeneity (typically diagonal)
- Further recall that the MAP prior is simply the prior induced by the Bayesian hierarchical model (BHM).
- This is **different** than the **robust** MAP prior, which is a **finite mixture model** consisting of the MAP and a vague prior.

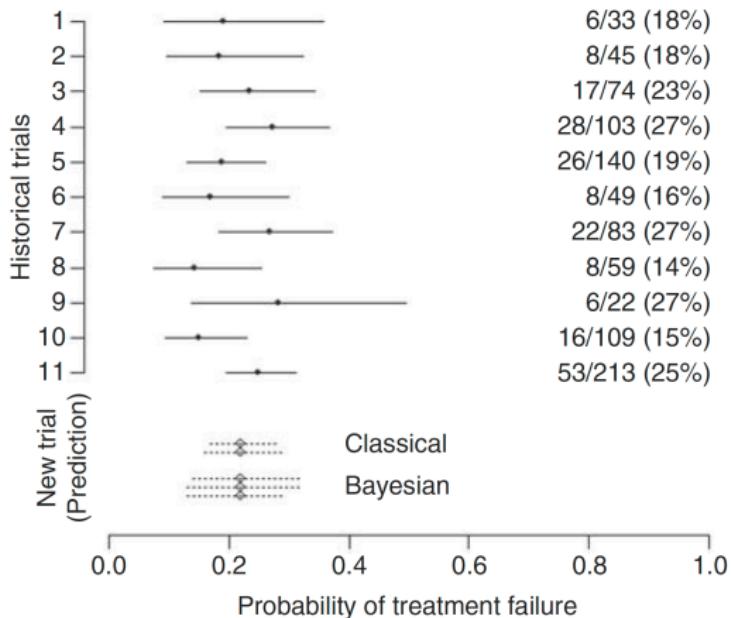
## Case Study 3: Meta-analytic predictive Prior

- Consider a phase IV trial in *de novo* transplant patients designed to compare a standard treatment (control) to a new intensified treatment of the same compound.
- Primary outcome: treatment failure (binary)
- A conventional balanced design requires  $n_j = 450$  subjects per arm,  $j = 0, 1$
- Due to study costs and time of recruitment, the clinical team hoped to reduce the number of control subjects in the trial by incorporating historical information.

## Case Study 3: Meta-analytic predictive Prior

- $K = 11$  historical studies were available of essentially identical design.
- The total number of historical controls available is  $\sum_{k=1}^{11} n_{0k0} = 930$
- The observed proportions of treatment failures range from 14% to 27%.
- The smallest and largest studies have sample sizes of  $n_{0(1)0} = 22$  and  $n_{0(K)0} = 213$ , with proportions  $\hat{p}_{0(1)0} = 27\%$ ,  $\hat{p}_{0(K)0} = 25\%$ .

## Case Study 3: Meta-analytic predictive Prior

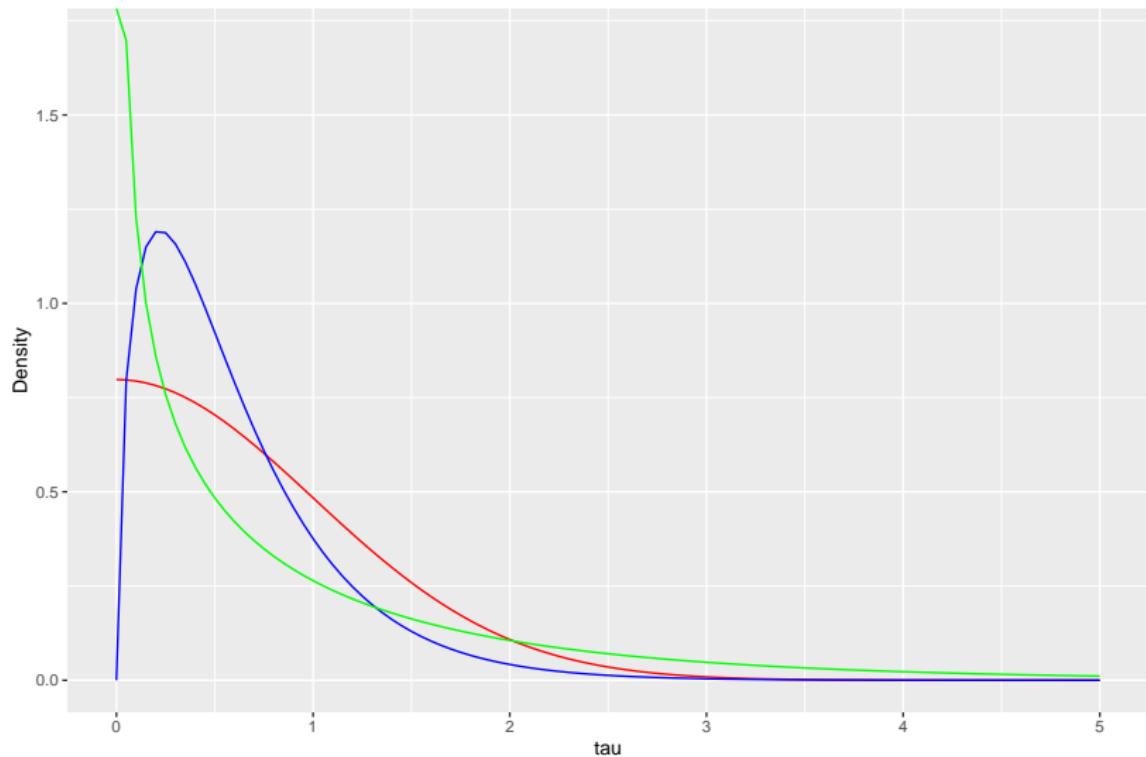


Observed proportions of treatment failures with 95% confidence intervals for 11 historical trials and predictions derived from meta-analytic-predictive approaches.

## Case Study 3: Meta-analytic predictive Prior

- For the between-trial standard deviation  $\tau$ , the three priors used were
  - $\tau \sim N^+(0, 1)$ ,
  - $\tau \sim \text{Gamma}(1.571, 2.590)$
  - $\tau \sim \text{Gamma}(0.577, 0.624)$
- First gamma prior was chosen on the basis that very small and large between trial heterogeneity are fairly unlikely (0.10 prior probability each).
- Second gamma prior was chosen to assess sensitivity allowing for higher probabilities of these two extreme scenarios (0.25 probability each).
- Sensitivity analyses were conducted by parameterizing by both the log-odds and the probability of failure.

## Case Study 3: Meta-analytic predictive Prior



Half-Normal( $0, 1$ ); Gamma( $1.571, 2.590$ ); Gamma( $0.577, 0.624$ )

## Case Study 3: Meta-analytic predictive Prior

	$\tau$ EST	95% CI	Log-odds $\theta^*$ EST	(SD)	$p^*$ EST	(95% CI)	$n^*$
Classical meta-analyses							
Pooled DL	0 0.12		-1.27 -1.28	(0.080) (0.166)	0.22 0.22	(0.19, 0.25) (0.17, 0.28)	930 275 <sup>n</sup> 220 <sup>a</sup> 214 <sup>b</sup> 196 <sup>c</sup>
PM	0.11		-1.28	(0.160)	0.22	(0.17, 0.28)	297 238 231 212
ML	0.11	(0.00, 0.39) <sup>p</sup>	-1.28	(0.156)	0.22	(0.17, 0.28)	313 250 243 234
REML	0.14		-1.29	(0.194)	0.22	(0.16, 0.29)	201 161 157 144
Bayesian meta-analyses							
Pooled HN <sub>sd=1</sub> Ga <sub>1,571,2,590</sub> Ga <sub>0,577,0,624</sub>	0 0.17 0.18 0.10	(0.01, 0.50) (0.02, 0.48) (0.00, 0.42)	-1.27 -1.29 -1.29 -1.29	(0.079) (0.253) (0.257) (0.202)	0.22 0.22 0.22 0.22	(0.19, 0.25) (0.14, 0.32) (0.14, 0.32) (0.15, 0.29)	930 90 90 146

Meta-analysis of 11 historical trials (Application 1): estimate of between-trial variation  $\tau$ , prediction of log-odds  $\theta^*$  and proportion  $p^*$ , and prior effective sample size  $n^*$  (Bayesian estimate = median).

## Case Study 4: Robust Meta-analytic predictive Prior

## Case Study 4: Robust Meta-analytic predictive Prior

- The previous example shows that, even under various levels of informativeness for  $\tau$ , the posterior results can be very similar.
- This makes the level of informativeness in the MAP prior difficult to control.
- To combat this, the **robust MAP** (rMAP) prior was developed.
- Recall that the rMAP prior is a conglomeration of a robust mixture prior and a MAP prior, i.e.,

$$\pi_{\text{rMAP}}(\theta|w) = (1 - w)\pi_{\text{MAP}}(\theta) + w\pi_{\text{vague}}(\theta), \quad w \in (0, 1).$$

- The value of  $w$  is elicited to control for the amount of borrowing.

## Case Study 4: Robust Meta-analytic predictive Prior

- Consider a proof-of-concept study in ulcerative colitis.
- Primary outcome: remission after 8 weeks of treatment (binary).
- $K = 4$  relevant historical placebo-controlled trials with a total of  $n_{00} = \sum_{k=1}^K n_{0k0} = 363$  placebo patients were identified, with remission rates in placebo ranging from 5.7% to 14.9%
- Based on the historical placebo data, the MAP prior for the remission rate in a new trial was derived. If  $\theta_0$  denotes log odds of responding under placebo,

$$\theta_0, \theta_{0k} | \sim N(\mu, \tau^2),$$

where  $\pi(\mu)$  is taken to be noninformative (but proper). The authors do not specify the prior on  $\mu$ .

- A weakly informative prior was given as  $\tau \sim N^+(0, 1)$ , which puts  $\Pr(\tau > 2) = 0.05$ .
- A value of  $\tau = 2$  corresponds to very large between-trial variability on the log-odds-scale, and would essentially lead to no borrowing from the historical data.

## Case Study 4: Robust Meta-analytic predictive Prior

- Note that a mixture prior requires **normalized** densities.
- It is not possible to normalize the MAP for non-normal models.
- The authors recommend a **mixture of conjugate priors**, e.g.,
  - ① Obtain samples of  $(\theta_{01}, \dots, \theta_{0K}, \mu, \tau)$  from the MAP.
  - ② Sample  $\theta_0 \sim N(\mu, \tau)$  (using the predictive distribution of the BHM).
  - ③ Convert the samples to the probability scale via  $p_* = (1 + \exp\{-\theta_0\})^{-1}$ .
  - ④ Approximate the induced prior on  $p_*$  via a mixture of  $L$  conjugate (beta) densities:

$$\pi_{\text{MAP}}(p_*) \approx \sum_{l=1}^L \gamma_l \text{Beta}(p_* | \alpha_{0l}, \beta_{0l}), \quad \gamma_l \in [0, 1], \quad \sum_{l=1}^L \gamma_l = 1$$

- The parameters for the approximation can be estimated by fitting a mixture model to the  $p_*$  samples using MLE (e.g., EM algorithm).

## Case Study 4: Robust Meta-analytic predictive Prior

- A mixture of conjugate priors results in a mixture of conjugate posteriors.
- E.g., for a Bernoulli model with a mixture prior consisting of  $L$  components,

$$p(p_* | \mathbf{y}) = \sum_{l=1}^L \tilde{\gamma}_l \text{Beta}(p_* | \tilde{\alpha}_{0l}, \tilde{\beta}_{0l}),$$

where

①  $\tilde{\alpha}_{0l} = y_0 + \alpha_{0l}$

②  $\tilde{\beta}_{0l} = n_0 - y_0 + \beta_{0l}$

③  $\tilde{\gamma}_l = \frac{\gamma_l B(\tilde{\alpha}_{0l}, \tilde{\beta}_{0l}) / B(\alpha_{0l}, \beta_{0l})}{\sum_{m=1}^L \gamma_m B(\tilde{\alpha}_{0m}, \tilde{\beta}_{0m}) / B(\alpha_{0m}, \beta_{0m})}$  is the new mixture weight (which depends on the data).

- Note: If using Stan, there is no benefit of using a conjugate prior since Stan does not exploit conjugacy.

## Case Study 4: Robust Meta-analytic predictive Prior

- Goal of the approximation is to minimize the KL divergence, given by

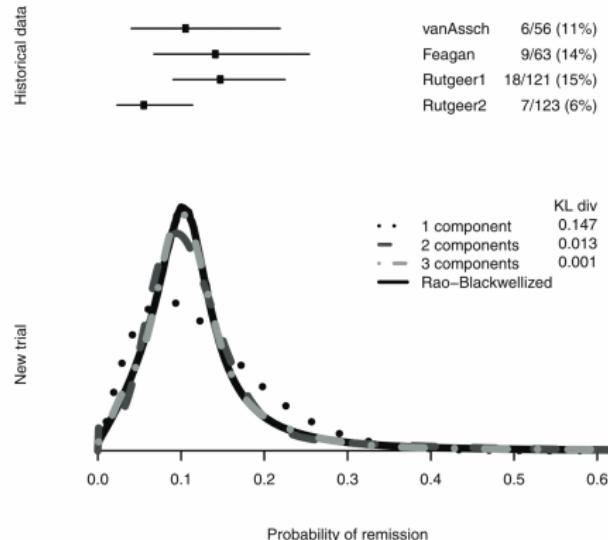
$$\text{KL}(\pi_{\text{MAP}}, \pi_{\text{approx}}) = \int \log \left( \frac{\pi_{\text{MAP}}(p_*)}{\pi_{\text{approx}}(p_*)} \right) \pi_{\text{MAP}}(p_*) dp_* . \quad (9)$$

- Suppose we possess samples  $\{p_*^{(m)}, m = 1, \dots, M\}$  from the MAP prior (transformed to the probability scale).
- The integral in (9) may be approximated via

$$\widehat{\text{KL}} = -\frac{1}{M} \sum_{m=1}^M \left( \log \pi_{\text{MAP}}(p_*^{(m)}) - \log \pi_{\text{approx}}(p_*^{(m)}) \right) . \quad (10)$$

- Since the first term in (10) does not depend on the parameters of the approximation, minimizing  $\widehat{\text{KL}}$  is equivalent to maximizing  $\log \pi_{\text{approx}}(p_*^{(m)}) \rightarrow \text{MLE}$  treating  $p_*^{(m)}$ 's as data!

## Case Study 4: Robust Meta-analytic predictive Prior



Observed placebo remission rates from four historical ulcerative colitis trials with 95% intervals, MAP prior for the rate in a new trial, and Beta mixtures with corresponding KL divergence

## Case Study 4: Robust Meta-analytic predictive Prior

- The mixture of three beta densities, which yields the best fit, is given by

$$\begin{aligned}\pi_{\text{approx}}(p_*) = & 0.53 \times \text{Beta}(p_*|2.5, 19.1) + 0.38 \times \text{Beta}(p_*|14.6, 120.2) \\ & + 0.08 \times \text{Beta}(p_*|0.9, 2.8).\end{aligned}$$

- Since this third component is not informative, the prior is already somewhat “robust.”
- Further robustification can be achieved via the **mixture**

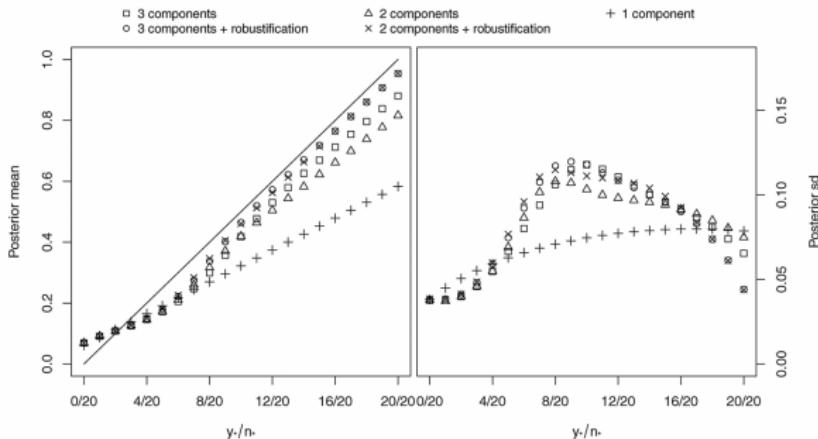
$$\pi(p_*) = (1 - \eta)\pi_{\text{approx}}(p_*) + \eta\pi_{\text{vague}}(p_*),$$

where we take  $\pi_{\text{vague}}(p_*) \propto 1$  and  $\eta = 0.1$ .

## Case Study 4: Robust Meta-analytic predictive Prior

- We now consider a new trial with  $n_* = 20$  placebo patients.
- The ESS of the three component approximation is 81. The ESS of the robustification is 63.
- We plot the posterior means and posterior SDs of  $p_*$  for different observed values of  $\hat{p}_0 = \frac{1}{20} \sum_{i=1}^{20} y_i$

## Case Study 4: Robust Meta-analytic predictive Prior



Posterior means and standard deviations (SD) versus observed placebo remission rates for all possible number of responders  $y^*$  with  $n^* = 20$  placebo patients, for different mixture priors approximating the MAP prior, and their robustifications.

## Case Study 4: Robust Meta-analytic predictive Prior

- For good agreement between new and historical data ( $y_0/20 \leq 4/20$ , say), the posterior means and standard deviations are very similar.
- For the range of **potential** prior-data conflict (5/20 to 10/20), uncertainty is increased for the mixture priors.
- Under high prior-data conflict ( $y_0/20 \geq 11/20$ ), the posterior weights discount the informative parts of the prior, and the posterior means are close to the observed rates.

## Case Study 4: Robust Meta-analytic predictive Prior

	$w_1$	$w_2$	$w_3$	$w_4$	Mean	2.5%	97.5%	ESS	Prior pred. prob. (%)
Prior $\hat{p}_H$	0.53	0.38	0.08		0.12	0.02	0.35	81	
Posterior for $y_*/n_*$									
0/20	0.62	0.30	0.08		0.07	0.01	0.15	78	14.9
2/20	0.50	0.46	0.04		0.11	0.04	0.20	110	59.6
5/20	0.59	0.31	0.11		0.17	0.08	0.33	74	13.7
10/20	0.25	0.01	0.74		0.42	0.20	0.64	14	1.5
15/20	0.004	0.00	0.996		0.67	0.47	0.84	24	0.3
Prior $\hat{p}_{HR}$	0.48	0.34	0.07	0.10	0.16	0.02	0.76	63	
Posterior for $y_*/n_*$									
0/20	0.60	0.29	0.08	0.03	0.07	0.01	0.15	76	13.9
2/20	0.49	0.45	0.04	0.02	0.11	0.04	0.21	108	55.1
5/20	0.54	0.28	0.10	0.08	0.18	0.08	0.37	69	20.0
10/20	0.11	0.00	0.32	0.56	0.46	0.23	0.69	20	6.6
15/20	0.00	0.00	0.16	0.84	0.72	0.51	0.88	22	3.1

Prior and posterior summaries for hypothetical trial outcomes  $y_*/n_*$ , for the three component mixture prior and its robustified version. The prior  $\hat{p}_H$  refers to the mixture approximation, and the prior  $\hat{p}_{HR}$  is the robustified version.

## Case Study 4: Robust Meta-analytic predictive Prior

	$w_1$	$w_2$	$w_3$	$w_4$	Mean	2.5%	97.5%	ESS	Prior pred. prob. (%)
Prior $\hat{p}_H$	0.53	0.38	0.08		0.12	0.02	0.35	81	
Posterior for $y_*/n_*$									
0/20	0.62	0.30	0.08		0.07	0.01	0.15	78	14.9
2/20	0.50	0.46	0.04		0.11	0.04	0.20	110	59.6
5/20	0.59	0.31	0.11		0.17	0.08	0.33	74	13.7
10/20	0.25	0.01	0.74		0.42	0.20	0.64	14	1.5
15/20	0.004	0.00	0.996		0.67	0.47	0.84	24	0.3
Prior $\hat{p}_{HR}$	0.48	0.34	0.07	0.10	0.16	0.02	0.76	63	
Posterior for $y_*/n_*$									
0/20	0.60	0.29	0.08	0.03	0.07	0.01	0.15	76	13.9
2/20	0.49	0.45	0.04	0.02	0.11	0.04	0.21	108	55.1
5/20	0.54	0.28	0.10	0.08	0.18	0.08	0.37	69	20.0
10/20	0.11	0.00	0.32	0.56	0.46	0.23	0.69	20	6.6
15/20	0.00	0.00	0.16	0.84	0.72	0.51	0.88	22	3.1

Prior and posterior summaries for hypothetical trial outcomes  $y_*/n_*$ , for the three component mixture prior and its robustified version

## Case Study 5: The Scale Transformed Power Prior

## Case Study 5: The Scale Transformed Power Prior

- The Comprehensive Post-Acute Stroke Services (COMPASS) study was a two-arm, cluster-randomized pragmatic trial designed to evaluate the effectiveness of a novel transitional care model (COMPASS care model) compared to usual care in mild-to-moderate stroke and transient ischemic attack (TIA) patients across a diverse set of hospitals within North Carolina, USA.
- The study consisted of two phases.
  - ① In Phase 1 of the COMPASS study, 40 hospital units were randomized in a 1:1 allocation scheme to either implement the COMPASS care model (i.e., the intervention) or to maintain their usual care practices.
  - ② In a second phase (Phase 2; an optional extension phase), intervention hospitals attempted to sustain real-world delivery of the intervention with minimal support.

## Case Study 5: The Scale Transformed Power Prior

- Of note, Phase 2 of the COMPASS study added a continuous measure of physical health (the PROMIS Global Health Scale) that was not collected in Phase 1.
- We consider the analysis of Phase 2 patient outcomes based on the PROMIS measure from one large hospital that provided the COMPASS care model during both phases of the study.
- Since the PROMIS outcome was not collected for Phase 1 patients, we consider the incidence of one or more falls as the Phase 1 outcome.
  - ▶ This variable is an indicator of whether the participant had at least one fall between hospital discharge and 90 days post-stroke (no falls versus at least one fall).
- Note that the historical and current outcomes measure related concepts (global disability versus global health) but are different scales (e.g., one binary, one continuous).

## Case Study 5: The Scale Transformed Power Prior

- To summarize,

Phase 1 (historical)	Incidence of falls	Binary
Phase 2 (current)	PROMIS score	Continuous

- The analysis assumes that the historical patient outcomes are independently distributed according to a logistic regression model and the current patient outcomes are independently distributed according to a linear regression model.

## Case Study 5: The Scale Transformed Power Prior

- The covariates of interest for our analyses were
  - ① An indicator for receipt of the eCare plan within 30 days of hospital discharge
  - ② An indicator for having a history of stroke or TIA
  - ③ An indicator for having non-white race
  - ④ Categorized NIH stroke scale score (NIHSS; 0 = no stroke symptoms, 1-4 = minor symptoms, and  $\geq 5$  = moderate-to-severe symptoms)
- As the incidence of falls outcome was collected during Phase 2 of the COMPASS study, a simple logistic regression model with incidence of falls as the outcome and the continuous PROMIS measure as the covariate was fit.
- We estimated the area under the Receiver Operating Characteristic (ROC) curve to be 0.64 indicating fair predictive ability of the incidence of falls for the PROMIS measure.

## Case Study 5: The Scale Transformed Power Prior

- Recall that the straPP assumption is

$$l_1(\theta_1|D_0)^{1/2}\theta_1 = l_0(\theta_0|D_0)^{1/2}\theta_0, \quad (11)$$

where 1 indexes the current data parameters and 0 indexes the historical parameters.

- Assume  $\sigma^2$  is known and fixed. In this case,  $\theta_1 = \beta_1$
- $l_1(\theta_1|D_0) = \mathbf{X}'_0 \mathbf{X}_0 / \sigma^2$  for the normal linear model, hence, we can solve for  $\theta_1$  in (11)

$$\beta_1 = \sigma [\mathbf{X}'_0 \mathbf{X}_0]^{-1/2} [\mathbf{X}'_0 W_0(\beta_0|\mathbf{X}_0) \mathbf{X}_0]^{1/2} \beta_0,$$

where  $W_0(\beta_0|\mathbf{X}_0) = \text{diag} \left\{ \frac{e^{x'_0 \beta_0}}{1 + e^{x'_0 \beta_0}}, i = 1, \dots, n_0 \right\}$ .

## Case Study 5: The Scale Transformed Power Prior

- When the straPP assumption in (11) does not hold, we may use the generalized straPP (gen-straPP), which assumes that the transformation differs by a location parameter  $\mathbf{c}_0$ , i.e.,

$$l_1(\boldsymbol{\theta}_1|D_0)^{1/2}\boldsymbol{\theta}_1 = l_0(\boldsymbol{\theta}_0|D_0)^{1/2}\boldsymbol{\theta}_0 + \mathbf{c}_0 \quad (12)$$

- We may similarly solve for  $\boldsymbol{\theta}_1 = \boldsymbol{\beta}_1$  via

$$\boldsymbol{\beta}_1 = \sigma [\mathbf{X}'_0 \mathbf{X}_0]^{-1/2} \left\{ [\mathbf{X}'_0 W_0(\boldsymbol{\beta}_0|\mathbf{X}_0) \mathbf{X}_0]^{1/2} \boldsymbol{\beta}_0 + \mathbf{c}_0 \right\}$$

- We assume  $\mathbf{c}_0 \sim N_p(0, \tau^2 \mathbf{I}_p)$  and set a prior on  $\tau$ , e.g.,  $\tau \sim N^+(0, 1)$ .

## Case Study 5: The Scale Transformed Power Prior

- Recall that both the straPP and the gen-straPP assume a power prior for the historical data, i.e.,

$$\pi(\beta_0 | D_0, a_0) \propto \left[ \prod_{i=1}^{n_0} p_{0i}^{y_{0i}} (1 - p_{0i})^{1-y_{0i}} \right] \pi_0(\beta_0),$$

where  $p_{0i} = \frac{e^{x'_{0i}\beta_0}}{1+e^{x'_{0i}\beta_0}}$ .

- For posterior inference, we conduct sampling on the  $\beta_0$  scale since we can solve  $\beta_1 = g(\beta_0 | \mathbf{c}_0)$ .
- The posterior from which we sample is thus

$$p(\beta_0, \sigma^2 | \mathbf{c}_0, D, D_0) \propto \mathcal{L}(g(\beta_0 | \mathbf{c}_0), \sigma^2 | D) \mathcal{L}(\beta_0 | D_0)^{a_0} \pi_0(\beta_0) \pi(\sigma^2)$$

## Case Study 5: The Scale Transformed Power Prior

- In practice,  $\sigma^2$  is unknown, but we use **partial borrowing** techniques to avoid having the historical data depend on the variance.
- Since the data are measured on different scales, we also use **partial borrowing** to not borrow from the intercept term.
- Thus, we only borrow from the treatment effect and the covariates.
- For comparison purposes, we fit the straPP, gen-straPP, power prior, and commensurate prior to the COMPASS data.
- To compare the overall quality of models fit based on the set of selected priors, we used the deviance information criterion (DIC)[17].

## Case Study 5: The Scale Transformed Power Prior

Model	$a_0$				
	0.10	0.25	0.50	0.75	1.00
Gen-straPP	2815.38	2815.65	2816.38	2816.82	2816.93
straPP	2815.38	2815.37	2817.25	2819.23	2821.30
PP	2816.44	2819.01	2822.20	2823.83	2824.92

DIC for the Gen-straPP, straPP and PP with Various Values of  $a_0$

# Case Study 5: The Scale Transformed Power Prior

Model	$a_0$	DIC	eCare Plan			History of Stroke			Minor NIHSS			Moderate-Severe NIHSS			Non-white			
			Mean	(SD)	RPV	Mean	(SD)	RPV	95% HPD	Mean	(SD)	RPV	95% HPD	Mean	(SD)	RPV	95% HPD	
Gen-straPP	0.10	2815.38	0.80	(0.92)	1.17	(-1.01, 2.60)	-0.97	(1.13)	1.17	(-3.13, 1.30)	-1.29	(1.05)	1.08	(-3.34, 0.78)	-3.54	(1.16)	1.48	(-5.85, -1.35)
straPP	0.25	2815.37	0.47	(0.85)	1.00	(-1.21, 2.14)	-0.54	(1.05)	1.00	(-2.52, 1.56)	-1.05	(1.01)	1.00	(-3.04, 0.92)	-2.79	(0.95)	1.00	(-4.63, -0.89)
RP	-	2816.65	1.14	(0.97)	1.29	(-0.78, 3.02)	-1.05	(1.22)	1.36	(-3.43, 1.36)	-1.29	(1.10)	1.18	(-3.46, 0.87)	-4.27	(1.31)	1.89	(-6.83, -1.70)
PP	0.10	2816.44	0.46	(0.76)	0.80	(-1.04, 1.96)	-0.25	(0.96)	0.83	(-2.17, 1.57)	-0.53	(0.80)	0.63	(-2.12, 1.02)	-2.80	(1.17)	1.51	(-5.14, -0.58)
COM	-	2818.47	0.53	(0.84)	0.97	(-1.01, 2.28)	-0.31	(0.96)	0.83	(-2.32, 1.41)	-0.25	(0.86)	0.73	(-2.04, 1.36)	-2.01	(1.25)	1.72	(-4.55, 0.23)
															-1.45	(1.34)	1.47	(-4.23, 1.08)

## Posterior Estimates for the COMPASS Data

RPV, ratio of posterior variances; Gen-straPP, generalized scale transformed power prior; straPP, scale transformed power prior; RP, reference prior; PP, power prior; COM, commensurate prior.

## Case Study 5: The Scale Transformed Power Prior

- The analyses with the Gen-straPP and straPP resulted in the smallest DICs when compared to analyses with other priors.
  - ▶ This suggests that the rescaling action of the straPP family is useful for translating the information on covariate effects from the incidence of falls outcome to the continuous PROMIS outcome.
- Aside from the general performance of the priors as measured by DIC, we also investigated the posterior estimates for the eCare Plan effect of interest.
  - ▶ Compared to the straPP, power and commensurate priors, the posterior mean effect based on the Gen-straPP is much closer to the value based on the reference prior.
  - ▶ While the posterior variance is reduced for analysis based on the straPP family of priors compared to the reference prior, the degree of variance reduction is substantially less than that based on the power and commensurate priors.

## Case Study 6: Propensity score integrated power prior

## Case Study 6: Propensity score integrated power prior

- An investigational device was newly developed to treat a coronary artery disease (CAD), which had been commonly treated with an approved medical device, considered as an active control.
- The plan was to conduct a randomized clinical trial in which the control arm was to be augmented by two identified external data sources:
  - ① Registry database ( $n_{20} = 1053$  fitting I/E Criteria).
  - ② Historical clinical study ( $n_{30} = 350$  fitting I/E Criteria).
- The registry database was recently completed, while the historical study was older.

## Case Study 6: Propensity score integrated power prior

- The primary endpoint was the occurrence of any specified adverse events at one year.
- The null and alternative hypotheses were

$$H_0 : \mu \geq 0 \text{ versus } H_1 : \mu < 0,$$

where  $\mu = \theta^{(1)} - \theta^{(0)}$  is the difference in one-year adverse event rate between the investigational device and the control.

- We reject the null hypothesis if

$$\Pr(\mu < 0 | D) \geq 0.975,$$

asymptotically assuring a type I error rate of 0.025 (one-sided).

## Case Study 6: Propensity score integrated power prior

- Let  $n_{jk}$  denote the number of patients assigned to arm  $k \in \{0, 1\}$  of study  $j \in \{C, R, H\}$ , where
  - $j = C$  denotes the **C**urrent RCT
  - $j = R$  denotes the **R**egistry
  - $j = H$  denotes the **H**istorical RCT
- For sample size calculations, the true values were taken to be  $\mu_0 = 0.192$  and  $\mu_1 = 0.12$ .
- This results in a sample size of  $n_{10}^* = n_{11}^* = 400$  per group to achieve 80% power.
- The RCT would enroll a total of 600 patients
  - $n_{C0} = 200$  patients were assigned to control.
  - $n_{C1} = 400$  patients were assigned to treatment.
- $A_R = 130$  subjects would be obtained from the registry.
- $A_H = 70$  subjects would be obtained from the historical trial.

## Case Study 6: Propensity score integrated power prior

- Denote the study ID as  $S_{Ci} = 1$  for all subjects in the current data set and  $S_{Ri} = 0$  and  $S_{Hi} = 0$  for all subjects in the historical data sets.
- Pooling all the data sources, a logistic regression model was fit to model the study ID as a function of baseline covariates, e.g.,

$$e_{ki} = \Pr(S_{ki} = 1 | \mathbf{X} = \mathbf{x}_{ki}; \boldsymbol{\beta}) = \frac{e^{\mathbf{x}'_{ki}\boldsymbol{\beta}}}{1 + e^{\mathbf{x}'_{ki}\boldsymbol{\beta}}},$$

for  $k \in \{C, R, H\}$  and  $i = 1, \dots, n_k$ .

- **Trimming step:** Subjects from  $k = R$  and  $k = H$  were excluded if their observed  $\hat{e}_{ki}$  was not in the range of the observed current data, i.e., they were excluded if

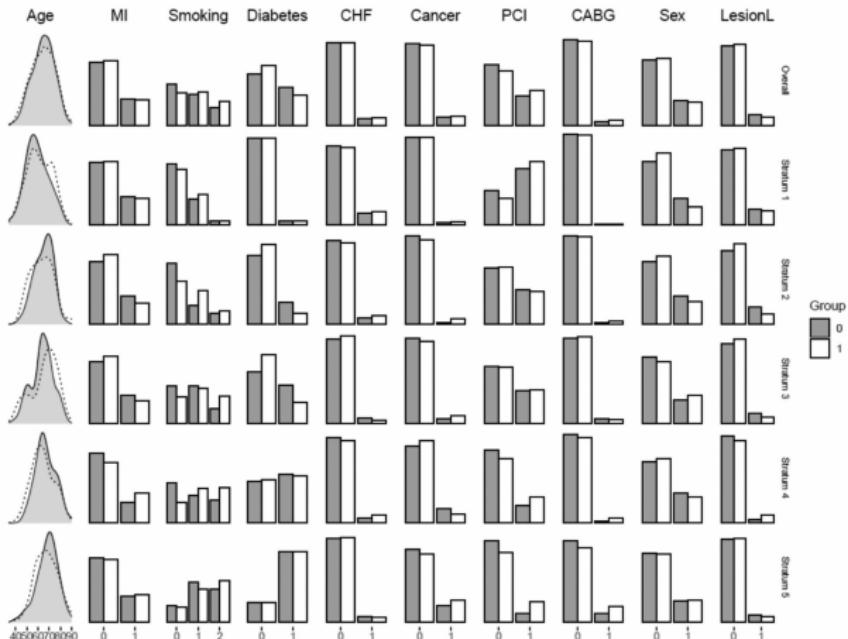
$$\hat{e}_{ki} \notin [e_{C(1)}, e_{C(n_C)}], \quad k \in \{R, H\},$$

where  $(j)$  denotes the  $j^{th}$  order statistic.

## Case Study 6: Propensity score integrated power prior

- After the trimming step, 1042 out of 1053 and 349 out of 350 subjects were retained from  $R$  and  $H$ , respectively.
- $M = 5$  strata were created based on the quintiles of the estimated propensity score for the **current data subjects** ( $K = C$ ).
- This results in approximately 120 subjects of the current data set in each stratum.
- The covariate distributions were examined post stratification to ensure overlap (i.e., balance) within each stratum.
- Balance could be assessed via standardized mean differences (SMDs), which is commonly used in the causal inference literature.

## Case Study 6: Propensity score integrated power prior



Balance checking of covariates between current study (group = 0) and historical clinical study (group = 1)

## Case Study 6: Propensity score integrated power prior

		Stratum					
		$k = 1$	$k = 2$	$k = 3$	$k = 4$	$k = 5$	Total
Current study ( $j = C$ )	$N_{k,C}$	120	120	121	119	120	600
Investigational device	$N_{k,C}^{(1)}$	79	83	87	71	84	404
Active control	$N_{k,C}^{(0)}$	41	37	34	48	36	196
Historical clinical study ( $j = H$ )	$N_{k,H}$	70	74	63	48	94	349
Overlapping coefficient	$r_{k,H}$	0.78	0.85	0.70	0.69	0.87	
	$r_{k,H}/\sum r_{k,H}$	0.20	0.22	0.18	0.18	0.22	
$\lambda_{k,H}$	$A_H r_{k,H}/\sum r_{k,H}$	13.96	15.33	12.68	12.38	15.65	70
$a_{k,H}$	$\lambda_{k,H}/N_{k,H}$	0.20	0.21	0.20	0.26	0.17	
Registry ( $j = R$ )	$N_{k,R}$	376	278	174	97	117	1042
Overlapping coefficient	$r_{k,R}$	0.82	0.78	0.79	0.74	0.79	
	$r_{k,R}/\sum r_{k,R}$	0.21	0.20	0.20	0.19	0.20	
$\lambda_{k,R}$	$A_R r_{k,R}/\sum r_{k,R}$	27.08	25.74	26.38	24.71	26.10	130
$a_{k,R}$	$\lambda_{k,R}/N_{k,R}$	0.07	0.09	0.15	0.25	0.22	

Summary of stratification.  $k$  denotes the stratum;  $\lambda_{kj}$  denotes the number of patients to be borrowed from stratum  $k$  in study  $j$ ,  $\alpha_{kj}$  denotes power prior parameter in study  $j$ ;  $r_{kj}$  denotes the overlapping coefficient (function of PS overlap).

## Case Study 6: Propensity score integrated power prior

- The final analysis was conducted after the clinical outcome data had been collected from all the 600 subjects in the current study.
- The posterior probability was computed as 99.8%, meeting the study success criterion.
- The treatment effect was computed as

$$\mu = \frac{1}{5} \sum_{k=1}^5 \left( \theta_k^{(1)} - \theta_k^{(0)} \right),$$

where  $\theta_k^{(j)}$  is the probability of having at least one adverse event in arm  $j$  of stratum  $k$ .

- Posterior means and credible intervals are displayed on the next slide.

## Case Study 6: Propensity score integrated power prior

	1	2	Stratum 3	4	5	Overall
	$\theta_1^{(1)}$	$\theta_2^{(1)}$	$\theta_3^{(1)}$	$\theta_4^{(1)}$	$\theta_5^{(1)}$	
Posterior mean	0.156	0.101	0.119	0.049	0.159	
Lower 95% credible interval	0.084	0.047	0.062	0.012	0.090	
Upper 95% credible interval	0.241	0.175	0.194	0.108	0.242	
	$\theta_1^{(0)}$	$\theta_2^{(0)}$	$\theta_3^{(0)}$	$\theta_4^{(0)}$	$\theta_5^{(0)}$	
Posterior mean	0.187	0.160	0.153	0.193	0.242	
Lower 95% credible interval	0.113	0.089	0.082	0.117	0.155	
Upper 95% credible interval	0.277	0.248	0.243	0.280	0.341	
	$\theta_1^{(1)} - \theta_1^{(0)}$	$\theta_2^{(1)} - \theta_2^{(0)}$	$\theta_3^{(1)} - \theta_3^{(0)}$	$\theta_4^{(1)} - \theta_4^{(0)}$	$\theta_5^{(1)} - \theta_5^{(0)}$	$\mu$
Posterior mean	-0.031	-0.059	-0.033	-0.144	-0.084	-0.070
Lower 95% credible interval	-0.147	-0.166	-0.141	-0.242	-0.204	-0.119
Upper 95% credible interval	0.082	0.043	0.070	-0.050	0.036	-0.023

Posterior means and credible intervals using the PSIPP.

## Case Study 7: Information Matrix Prior

## Case Study 7: Information Matrix Prior

- The misregulation of the chromatin structure in DNA is associated with the progression of cancer, aging, and developmental defects.
- It is known that the accessibility of genetic information in DNA is dependent on the positioning of histone proteins packaging the chromatin, forming nucleosomes
- Nucleosome positioning is known to be influenced by di- and tri-nucleotide repeats, but overall, the sequence signals influencing positioning are relatively weak and difficult to detect.
- We consider a genome-wide study of chromatin structure in yeast (Hogan, Lee, and Lieb (2006)).
- The data consist of normalized log-ratios of intensities measured for a tiled array for chromosome III, consisting of 50-mer oligonucleotide probes that overlap every 20 bp.

## Case Study 7: Information Matrix Prior

- A two-state Gaussian hidden Markov model (HMM) was fit to determine the nucleosomal state for each probe.
- The primary interest was to determine whether certain sequence features are predictive of nucleosome and nucleosome-free positions in the genome.

## Case Study 7: Information Matrix Prior

- Recall that the IM prior (without ridge) is given by

$$\pi_{\text{IM}}(\beta | \mu_0, c_0) \propto |\mathcal{I}(\beta)|^{1/2} \exp \left\{ -\frac{1}{2c_0} (\beta - \mu_0)' \mathcal{I}(\beta) (\beta - \mu_0) \right\}$$

- We concentrated on a region of about 1400 adjacent probes on yeast chromosome III.
- For each probe, the covariate vector was the set of observed frequencies of nucleotide  $k$ -tuples, with  $k = 1, 2, 3, 4$ . This led to a total of  $p = 340$  covariates + an intercept term.
- The HMM-based classification gave the observed “state” of each probe, whether corresponding to a nucleosome-free region (NFR) or a nucleosome(N).
- We fit the prior using  $c_0 \in \{1, 10\}$ , reporting only the former as the results were similar.

## Case Study 7: Information Matrix Prior

- We carried out ten-fold cross validation to test the predictive power of the model.
- The set of probes, with the associated covariates, were divided into ten non-overlapping pairs of training sets (90% of probes) and test sets (10% of probes).
- For each training-test set pair, we:
  - ① Fit the logistic regression model to the training data set,
  - ② Used the fitted values of  $\beta$  to compute the posterior probabilities of classification into the NFR state for the corresponding test set.
- The sensitivity and specificity of cross validation using the three different priors was compared, where any region having an estimated posterior probability of  $\geq 50\%$  with the logistic model was classified as an NFR.

## Case Study 7: Information Matrix Prior

Set	Prior	Range[ $E(\beta y)$ ]	%pN F R	ave.corr
(a)	$IMR$	(-0.295, 0.443)	0.708	0.664
	$N(0, I)$	(-0.192, 0.161)	0.123	0.509
	$N(0, 10^6 I)$	(-0.188, 0.168)	0.111	0.475
	$gBMA$	(-8.778, 11.437)	0.569	0.670

Overall sensitivity and specificity of methods using three types of priors, under the full model, and BMA using the g-prior (gBMA), by ten-fold cross validation, on set (a):  $p = 340$ ,  $n = 1260$ . %pN F R: percentage of predicted nucleosome-free regions by each method; ave.corr: average percent correct classification = (Sens + Spec)/2, averaged over the ten cross-validation data sets.

## Case Study 7: Information Matrix Prior

- Recall that the IMR prior is given by

$$\pi_{\text{IMR}}(\boldsymbol{\beta} | \boldsymbol{\mu}_0, c_0, \lambda_0) \propto |\mathcal{I}(\boldsymbol{\beta}) + \lambda_0 \mathbf{I}_p|^{1/2}$$

$$\exp \left\{ -\frac{1}{2c_0\sigma^2} (\boldsymbol{\beta} - \boldsymbol{\mu}_0)' [\mathcal{I}(\boldsymbol{\beta}) + \lambda_0 \mathbf{I}_p] (\boldsymbol{\beta} - \boldsymbol{\mu}_0) \right\}$$

- We increased the number of predictors to test how far the improved model fit, by including the 5-tuple counts, would be offset by the increased covariate dimensionality.
- The same process to forming training-test pairs was conducted, except  $k$ -tuples up to 5 yields  $p = 1364$  covariates, with the training sample size being  $n_{\text{train}} = 1260$ .
- The IMR prior in this case is the only method that could predict even a proportion of the NFRs correctly.
- However, the overall predictive power using 5-mers decreased, due to a combination of overfitting with sparse data as well as induced bias due to the massive increase in dimensionality.

## Case Study 7: Information Matrix Prior

Set	Prior	Range[ $E(\beta y)$ ]	%pNFR	ave.corr
(a)	$IMR$	(-0.295, 0.443)	0.708	0.664
	$N(0, I)$	(-0.192, 0.161)	0.123	0.509
	$N(0, 10^6 I)$	(-0.188, 0.168)	0.111	0.475
	$gBMA$	(-8.778, 11.437)	0.569	0.670
(b)	$IMR$	(-0.518, 0.535)	0.387	0.447
	$N(0, I)$	(-1.142, 0.496)	0	-
	$N(0, 10^6 I)$	(-3.424, 2.179)	0	-

Overall sensitivity and specificity of methods using three types of priors, under the full model, and BMA using the g-prior (gBMA), by ten-fold cross validation, on set (a):  $p = 340$ ,  $n = 1260$  and set (b):  $p = 1364$ ,  $n = 1260$ .

## Case Study 7: Information Matrix Prior

- The IMR prior showed uniformly higher sensitivity while its specificity was comparable to the other two priors.
- The IMR predicted a slightly higher percentage of NFRs than the true percentage, while the other two methods consistently underestimated the number of NFRs.
- Out of 340 covariates using the IMR prior, 93 and 91 coefficients had approximate 95% HPD intervals above and below zero.
- Among the significant dinucleotides, “aa”, “at”, “tg”, and “tt” had a positive effect on the possibility of being an NFR, while “ac”, “ag”, “cc”, “ct”, “gc”, and “gg” had the opposite effect.
- It was previously found that “aa” or “tt” repeats have an effect of making DNA rigid, and thus difficult to form nucleosomes, while “gg” and “cc” lead to less rigid DNA for which it is easier to form nucleosomes (Thastrom et al. (2004)).

## Case Study 8: Conjugate Prior for GLMs

## Case Study 8: Conjugate Prior for GLMs

- We consider the data from Finney (1947), obtained to study the effect of the rate and volume of air inspired on a transient vaso-constriction in the skin of the digits.
- The response variable measured is binary with 1 and 0 indicating occurrence or nonoccurrence of vaso-constriction, respectively.
- There are  $n = 39$  observations in the data set (note, asymptotics would be inappropriate here).
- The two covariates are  $x_1 = \log(\text{volume})$  and  $x_2 = \log(\text{rate})$  with  $\beta_1$  and  $\beta_2$  denoting the respective regression coefficients.

## Case Study 8: Conjugate Prior for GLMs

- We consider a logistic regression model. Let  $\mathbf{x}_i = (1, x_{1i}, x_{2i})'$  and  $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)'$ .
- The conjugate prior takes the form

$$\pi_{\text{CI}}(\boldsymbol{\beta}|a_0, \mathbf{y}_0) \propto \exp \left\{ a_0 \sum_{i=1}^n \left( y_{0i} \mathbf{x}'_i \boldsymbol{\beta} - \log \left( 1 + e^{\mathbf{x}'_i \boldsymbol{\beta}} \right) \right) \right\},$$

where  $y_{0i} = E_{\pi_{\text{CI}}} [y_i | \mathbf{x}_i] \in [0, 1]$  is the **prior prediction** for the mean of  $y_i$ .

# Case Study 8: Conjugate Prior for GLMs

## MLE of Finney Data

<i>Dependent variable:</i>	
	Resp
log Volume	5.220*** (1.858)
log Rate	4.631*** (1.789)
Intercept	-2.924** (1.288)
Observations	39
Log Likelihood	-14.632
Akaike Inf. Crit.	35.264

Note:

\* $p < 0.1$ ; \*\* $p < 0.05$ ; \*\*\* $p < 0.01$

## Case Study 8: Conjugate Prior for GLMs

- In order to demonstrate the sensitivity to the hyperparameter  $y_0$  and  $a_0$ , we analyze the posterior using several different values.
- In particular, we implement all combinations of  $y_{0i} \in \{0.1, 0.5\}$ ,  $a_0 \in \{1, 10, 100\}$

## Case Study 8: Conjugate Prior for GLMs

$y_0$	$a_0$	Parameter	Prior			Posterior		
			Mean	SD	95% HPD	Mean	SD	95% HPD
0.1	1	$\beta_0$	-2.700	0.888	(-4.328, -1.251)	-1.997	0.551	(-3.105, -0.974)
		$\beta_1$	0.067	1.262	(-2.471, 2.468)	1.850	0.649	( 0.600, 3.142)
		$\beta_2$	0.357	0.986	(-1.263, 2.237)	1.688	0.694	( 0.434, 3.098)
	10	$\beta_0$	-2.245	0.205	(-2.649, -1.846)	-2.071	0.194	(-2.456, -1.698)
		$\beta_1$	0.003	0.346	(-0.673, 0.687)	0.459	0.285	(-0.105, 1.013)
		$\beta_2$	0.039	0.232	(-0.398, 0.503)	0.344	0.233	(-0.095, 0.807)
0.5	100	$\beta_0$	-2.2019	0.063	(-2.322, -2.076)	-2.178	0.062	(-2.300, -2.058)
		$\beta_1$	-0.0002	0.108	(-0.210, 0.210)	0.061	0.106	(-0.147, 0.266)
		$\beta_2$	0.0039	0.070	(-0.134, 0.141)	0.040	0.070	(-0.096, 0.178)
	10	$\beta_0$	0.0014	0.406	(-0.804, 0.796)	-0.502	0.325	(-1.147, 0.123)
		$\beta_1$	0.0049	0.683	(-1.317, 1.368)	1.342	0.537	( 0.313, 2.415)
		$\beta_2$	-0.0025	0.482	(-0.979, 0.940)	0.897	0.413	( 0.119, 1.729)
0.5	100	$\beta_0$	0.0002	0.121	(-0.235, 0.236)	-0.071	0.114	(-0.301, 0.149)
		$\beta_1$	0.0003	0.207	(-0.415, 0.398)	0.213	0.198	(-0.165, 0.611)
		$\beta_2$	0.0003	0.135	(-0.269, 0.264)	0.127	0.130	(-0.132, 0.379)
	100	$\beta_0$	-0.0004	0.038	(-0.074, 0.074)	-0.007	0.038	(-0.079, 0.067)
		$\beta_1$	0.0006	0.065	(-0.129, 0.124)	0.023	0.065	(-0.104, 0.148)
		$\beta_2$	0.0006	0.042	(-0.083, 0.083)	0.013	0.042	(-0.070, 0.096)

Summary statistics from the prior and posterior distributions for Finney data.

## Case Study 9: The Hierarchical Prediction Prior

## Case Study 9: The Hierarchical Prediction Prior

- Recall that the HPP is an augmentation of the conjugate prior, where the prior prediction is treated as random. Specifically,

$$\pi_{\text{HPP}}(\beta, \mathbf{y}_0) \propto \frac{\exp \left\{ a_0 [\sum_{i=1}^n (y_{0i}\theta_i - b(\theta_i))] \right\}}{C(\mathbf{y}_0, a_0)} \pi_{\text{HPP}}(\mathbf{y}_0).$$

- The hyperprior can be any prior in the support of the mean.
- We again use the Finney data, and elicit a beta hyperprior, e.g.,

$$\pi_{\text{HPP}}(\mathbf{y}_0) \propto \prod_{i=1}^n y_{0i}^{\phi_{0i}\mu_{0i}-1} (1-y_{0i})^{\phi_{0i}(1-\mu_{0i})-1}$$

- This gives  $E(y_{0i}) = \mu_{0i}$  and  $\text{Var}(y_{0i}) = \frac{\mu_{0i}(1-\mu_{0i})}{1+\phi_{0i}}$ .

## Case Study 9: The Hierarchical Prediction Prior

- Suppose that an expert informs us that lower volume, on average, should yield a lower probability for the event, but is unsure of the effect of inhalation on the response.
- The expert then gives their prediction based on the level of volume, providing
  - ①  $\mu_{0i} = \mu_0 = 0.3$  when volume is less than 1
  - ②  $\mu_{0i} = \mu_1 = 0.7$  when volume is larger than 1
- The expert further dictates their uncertainty surrounding these values, which is captured by  $\phi_{0i} = \phi_0 > 0$ .
- The following hyperparameters for the exercise were used:
  - ▶  $(\mu_0, \mu_1) \in \{(0.3, 0.7), (0.7, 0.3)\}$
  - ▶  $\phi_0 \in \{1, 10\}$
  - ▶  $a_0 \in \{0.1, 0.5, 1.0\}$

# Case Study 9: The Hierarchical Prediction Prior

		$(\mu_0, \mu_1)$												
		(0.3, 0.7)				(0.7, 0.3)								
		Intercept		logVol		logRate		Intercept		logVol		logRate		
$a_0$	$\phi_0$	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd	
Prior	0.1	1	-0.370	2.294	3.681	4.536	0.117	3.025	0.344	2.278	-3.593	4.394	-0.092	3.045
		10	-0.261	2.080	3.245	3.862	0.007	2.730	0.308	2.067	-3.320	3.907	-0.063	2.681
	0.5	1	-0.199	0.793	1.896	1.533	0.085	1.011	0.192	0.788	-1.862	1.516	-0.062	0.992
		10	-0.186	0.666	1.745	1.256	0.079	0.822	0.191	0.665	-1.765	1.266	-0.073	0.821
	1	1	-0.190	0.562	1.691	1.116	0.089	0.704	0.186	0.564	-1.695	1.141	-0.079	0.709
		10	-0.175	0.447	1.588	0.863	0.071	0.541	0.187	0.457	-1.602	0.854	-0.090	0.545
Posterior	0.1	1	-2.357	1.061	4.740	1.556	3.781	1.466	-2.066	0.921	3.852	1.270	3.407	1.292
		10	-2.244	1.000	4.613	1.485	3.619	1.376	-1.990	0.895	3.743	1.244	3.298	1.248
	0.5	1	-1.262	0.646	3.263	1.020	2.071	0.872	-0.964	0.518	1.798	0.730	1.739	0.721
		10	-1.004	0.502	2.953	0.879	1.660	0.662	-0.804	0.451	1.499	0.657	1.494	0.619
	1	1	-0.959	0.520	2.823	0.858	1.553	0.699	-0.672	0.433	1.133	0.604	1.274	0.598
		10	-0.683	0.365	2.457	0.683	1.094	0.454	-0.471	0.341	0.707	0.515	0.955	0.459

Prior and posterior summary of the HPP using the Finney data.

## Case Study 9: The Hierarchical Prediction Prior

- Note that the precision can increase in one of two ways:
  - ➊ Increasing  $a_0$  (i.e., increasing the amount of borrowing).
  - ➋ Increasing  $\phi_0$  (i.e., higher precision in prior prediction).
- When the prior prediction was highly inaccurate (i.e., when  $(\mu_0, \mu_1) = (0.7, 0.3)$ ) the point estimates and standard errors were somewhat similar for a fixed value of  $a_0$ .
- Changing  $a_0$  has a comparatively larger impact on the posterior density.

## Case Study 10: PS Integrated Commensurate Prior

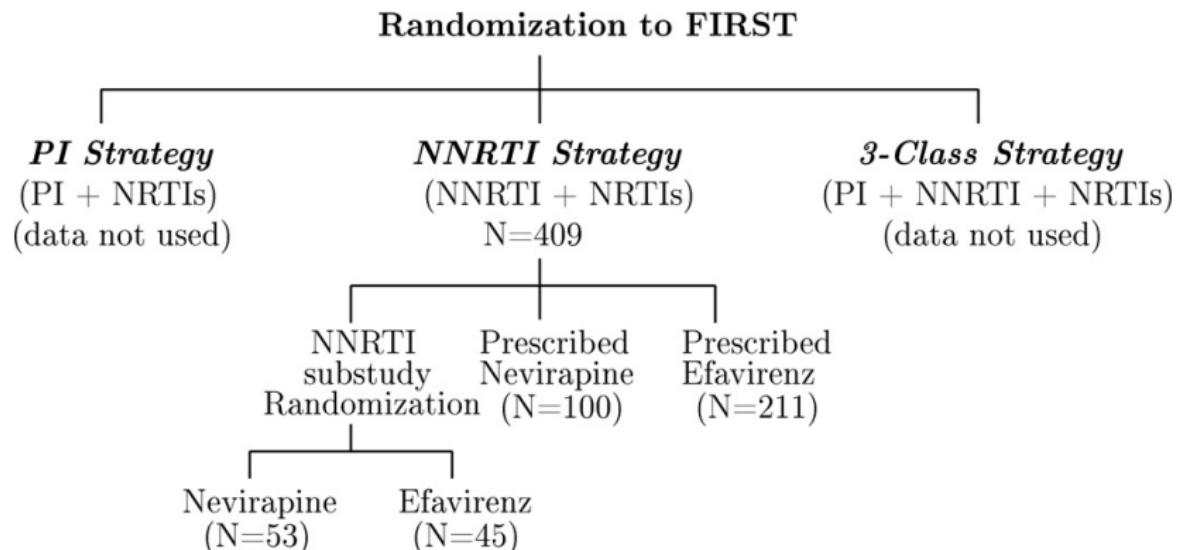
## Case Study 10: PS Integrated Commensurate Prior

- In some cases, statistical methods are needed to facilitate integrative analysis based on both RCT and observational study data.
- We consider the Flexible Initial Retrovirus Suppressive Therapies (FIRST) trial as an example.
- Highly active antiretroviral therapy-naive, HIV-infected subjects were randomized to three strategies (nucleoside reverse transcriptase inhibitor (NRTI) was used in all three strategies)
  - ① a two-class protease inhibitor (PI+NRTI)
  - ② a two-class non-nucleoside reverse transcriptase inhibitor (NNRTI+NRTI)
  - ③ a three-class strategy (PI+NNRTI+NRTI)

## Case Study 10: PS Integrated Commensurate Prior

- Participants within the two strategies involving NNRTIs could further specify whether they wanted to
  - ① be randomly assigned to a NNRTI drug (nevirapine, NVP, or efavirenz, EFV) before the randomization to strategy arms
  - ② permit a study clinician to prescribe one of the two drugs
- The three strategies were compared for long-term virological and immunological durability, drug resistance, and disease progression.

# Case Study 10: PS Integrated Commensurate Prior



Study design of the FIRST trial.

## Case Study 10: PS Integrated Commensurate Prior

- There are two common target estimands in the causal inference literature:
  - ① ATE = average treatment effect =  $E(Y_1 - Y_0)$ ,  $Y_j$  = potential outcome (outcome that *would have been observed*) if given treatment  $j$ .
  - ② ATT = average treatment effect for the treated =  $E(Y_1 - Y_0|Z = 1)$ , where  $Z$  = treatment indicator.
- In RCTs, since treatments are assigned at random, we assume that, on average, treated subjects have similar characteristics to those in the competing study arms. Thus, the ATT and ATE coincide.
- In observational studies it is necessary to assume that  $E(Y_1) \neq E(Y_1|Z = 1)$  (and similarly for the control group) due to selection bias.

## Case Study 10: PS Integrated Commensurate Prior

- Let  $e(\mathbf{x}) = \Pr(Z = 1 | \mathbf{X} = \mathbf{x})$  denote the propensity score (PS). It can be shown that

$$\mathbf{X} \perp \mathbf{Z} | e(\mathbf{X}),$$

i.e., the propensity score provides all the necessary information about treatment assignment.

- One can also show that

$$(Y_1 - Y_0) \perp \mathbf{Z} | \mathbf{X},$$

which implies that treatments may be viewed as being **randomly assigned** to subjects with roughly the same propensity score, yielding unbiased estimates of the ATE.

- This is the impetus for PS stratification—individuals with similar PS's can be treated as belonging to an RCT.

## Case Study 10: PS Integrated Commensurate Prior

- In this case study from Zhao et al. (2016) [21], we will use PS matching, which is shown to have superior reduction of selection bias compared to competitors.
- Recall that PS matching involves the following procedure:
  - ① Estimate the PS (e.g., logistic regression, random forests, etc.).
  - ② Use 1:1 nearest neighbor matching on the PS.
  - ③ Remove unmatched individuals from the data set.
- Typically, no participants will have exactly the same PS, so a caliper,  $c$ , is defined, e.g.,  $c = 0.01$ .
- The caliper allows individuals to be matched if  $|e(\mathbf{x}_i) - e(\mathbf{x}_j)| \leq c$

## Case Study 10: PS Integrated Commensurate Prior

	Randomized cohort			Non-randomized cohort		
	EFV (N=45)	NVP (N=53)	p-value	EFV (N=211)	NVP (N=100)	p-value
AGE	39.0±7.6	36.7±8.5	0.17	38.6±9.9	38.9±8.4	0.78
RACE (RACE=3)	12 (26.7)	14 (26.4)	0.98	61 (28.9)	26 (26.0)	0.59
podbl (=1)	16 (35.6)	22 (41.5)	0.55	83 (39.3)	38 (38.0)	0.82
cd4bl	227.7±207.3	209.5±193.0	0.66	190.2±189.0	242.9±227.3	<b>0.03</b>
lrnabb (log10(rnabl))	5.1±0.9	5.2±0.8	0.56	5.1±0.8	4.9±0.8	0.08
GENDER (=male)	35 (77.8)	40 (75.5)	0.79	167 (79.2)	77 (77.0)	0.67
malesex (=1)	21 (46.7)	22 (41.5)	0.61	100 (47.4)	41 (41.0)	0.29
idu (=1)	5 (11.1)	11 (21.2)	0.18	25 (11.9)	17 (17.0)	0.22

Baseline characteristics of the randomized and nonrandomized parts of the FIRST study. Race (white vs. others), progression of disease before randomization or not ("podbl", 1: yes), baseline average cd4 count ("cd4bl"), log baseline HIV RNA level ("lrnabb"), gender (1: male), malesex indicating homosexual activity for men (1: yes), and injection drug use ("idu", 1: yes). Continuous covariates show mean  $\pm$  SD. Binary covariates show N (%).

## Case Study 10: PS Integrated Commensurate Prior

- A Frequentist analysis of the data yielded a non-significant effect for the RCT but a significant effect for the randomized study (unadjusted for confounding), and the effects were of different signs.
  - ▶ Non-Randomized:  $-0.68$  95% CI : [1.25, 0.11].
  - ▶ Randomized:  $0.79$  95% CI : [0.54, 2.13].
- PS matching resulted in  $n_{NR}^* = 89$  matched pairs in the non-randomized study. Any unmatched individuals were discarded from the final analysis.
- A Frequentist analysis of the non-randomized study post matching yielded  $\hat{\lambda}_{NR}^* = 0.72$  with a 95% CI[1.43, 0.01].

## Case Study 10: PS Integrated Commensurate Prior

- Once the pairs have been matched, we may use Bayesian hierarchical modeling.
- Let the observations from the  $m^{th}$  matched pair be denoted as  $\mathbf{y} = (y_{m0}, y_{m1})'$ , where  $y_{m0}$  denotes the matched control and  $y_{m1}$  denotes the matched treated for  $m = 1, \dots, 89$ .
- We then assume the following model:

$$\text{logit} [\Pr(y_{mj} = 1)] = \alpha_m + \lambda_0 1\{j = 1\},$$

where  $\alpha_m$  is a match-specific intercept and  $\lambda_0$  is the treatment effect (log-odds ratio) for the nonrandomized study.

- We assume  $\alpha_m \sim N(\mu_0, \tau_1^{-1})$  with hyperpriors.

## Case Study 10: PS Integrated Commensurate Prior

- For the randomized study, we assume  $Y_i \sim Bernoulli(\pi_i)$  where  $\pi_i$  denotes the probability of VS, where

$$\text{logit}(\pi_i) = \mathbf{x}_i' \boldsymbol{\beta} + \lambda z_i,$$

and where  $z_i = 1$  if patient  $i$  received treated and 0 otherwise and  $\lambda$  is the log odds ratio.

- To incorporate the non-randomized study, a commensurate prior is specified for  $\lambda$ . Specifically,  $\lambda \sim N(\lambda_0, \tau_3^{-1})$ .
- A spike and slab prior for  $\tau_3$  is recommended, e.g.,

$$\tau_3 \sim p_0 \times 1\{\tau_3 = a\} + (1 - p_0) \times \text{Unif}(S_l, S_u),$$

where  $a > S_u$  and  $a$  is chosen to limit the amount of borrowing.

## Case Study 10: PS Integrated Commensurate Prior

- For the randomized study, we assume  $Y_i \sim Bernoulli(\pi_i)$  where  $\pi_i$  denotes the probability of VS, where

$$\text{logit}(\pi_i) = \mathbf{x}_i' \boldsymbol{\beta} + \lambda z_i,$$

and where  $z_i = 1$  if patient  $i$  received treated and 0 otherwise and  $\lambda$  is the log odds ratio.

- To incorporate the non-randomized study, a commensurate prior is specified for  $\lambda$ . Specifically,  $\lambda \sim N(\lambda_0, \tau_3^{-1})$ .
- A spike and slab prior for  $\tau_3$  is recommended, e.g.,

$$\tau_3 \sim p_0 \times 1\{\tau_3 = a\} + (1 - p_0) \times \text{Unif}(S_l, S_u),$$

where  $a > S_u$  and  $a$  is chosen to limit the amount of borrowing.

## Case Study 10: PS Integrated Commensurate Prior

- The hyperparameters  $a = 40$  and  $p_0 = 0.3$  were elicited.
- This prior in an effective sample size of 53 when the full sample size is 100 for the nonrandomized data.
- The final estimate of the log-odds ratio  $\lambda$  was  $-0.30$  with a 95% Bayesian credible interval (BCI) of  $[-0.88, 0.32]$ .
- The pooled treatment effect of  $-0.30$  lies in between those obtained from MLE analyses of the data sets alone.
- The CI width (1.20) is smaller than those obtained using the data sets individually (2.67 [rand] and 1.42 [non-rand]).
- The posterior mean of  $\tau_3$  was estimated as 25.85, showing a moderate degree of borrowing from the NR data.

## Case Study 11: Pediatric Extrapolation with Robust MAP

## Case Study 11: Pediatric Extrapolation with Robust MAP

- Systemic lupus erythematosus (SLE) is a relapsing, chronic, inflammatory autoimmune disease with diverse clinical and laboratory manifestations.
- Childhood-onset SLE (cSLE) is rare, with estimated annual incidence of 0.3 to 0.9 per 100,000 children.
- Compared with SLE starting in adulthood, there is higher disease activity; increased rates of renal, neurological and haematological involvement; and faster damage accrual over time with cSLE.

## Case Study 11: Pediatric Extrapolation with Robust MAP

- Despite intense efforts in recruitment, only 93 subjects were enrolled in the study, which was not enough to be adequately powered, and no formal statistical hypothesis testing was planned in the protocol.
- The clinical review proposed Bayesian methods as a means to borrow information from the adult to the pediatric population, expecting similarity of disease and response in these two populations.
- The method applied was a Bayesian mixture model with an informative prior based on a weighted combination of a skeptical prior with a mean effect size of zero and a meta-analytical prior from two adult studies.

## Case Study 11: Pediatric Extrapolation with Robust MAP

- Despite intense efforts in recruitment, only 93 subjects were enrolled in the study, which was not enough to be adequately powered, and no formal statistical hypothesis testing was planned in the protocol.
- The primary endpoint was the proportion of patients who met the SLE Responder Index (SRI) Response criteria at Week 52.
- The primary analysis failed to demonstrate a statistically significant difference between belimumab and placebo.

## Case Study 11: Pediatric Extrapolation with Robust MAP

	<b>Placebo <i>N</i> = 40<sup>a</sup></b>	<b>Belimumab 10 mg/kg <i>N</i> = 53</b>
Response, % ( <i>n</i> )	44% (17)	53% (28)
Observed difference	-	9.2%
Odds ratio (95% CI)	-	1.5 (0.6, 3.5)

Primary analysis of pediatric belimumab study.

## Case Study 11: Pediatric Extrapolation with Robust MAP

- The clinical review proposed Bayesian methods as a means to borrow information from the adult to the pediatric population.
- To conduct such an analysis, a prior distribution had to be determined.
- First, they identified the relevant data that could be used; specifically, there were two adult efficacy studies that compared two dose levels (1 mg/kg and 10 mg/kg) against placebo.
- These data were used since the clinical review team believed the disease and patient response to treatment were likely to be similar between the adult and pediatric subjects, given that the pediatric and adult diseases have similar underlying pathophysiology and management.

## Case Study 11: Pediatric Extrapolation with Robust MAP

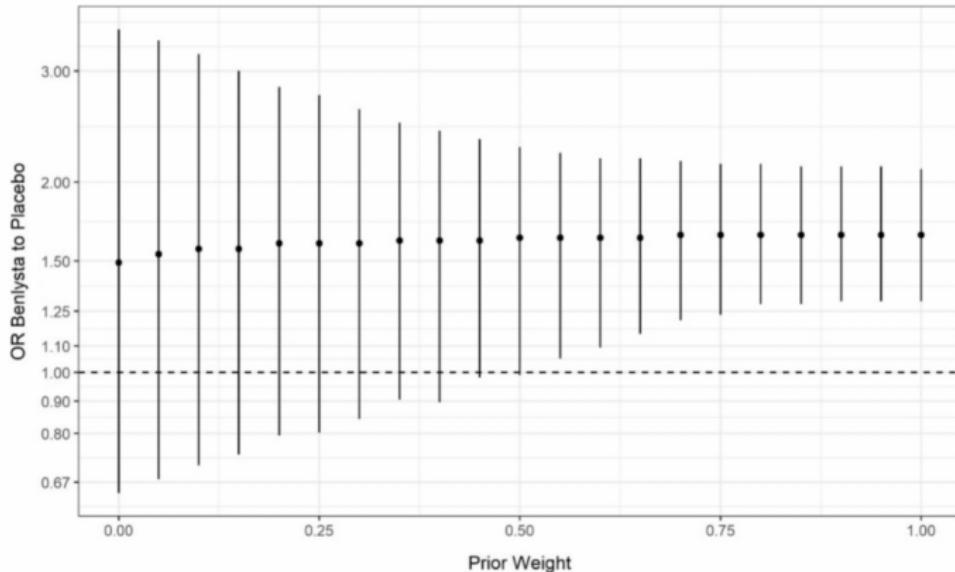
	Adult study 1		Adult study 2	
	Placebo <i>N</i> = 275	Belimumab 10 mg/kg <i>N</i> = 273	Placebo <i>N</i> = 287	Belimumab 10 mg/kg <i>N</i> = 290
Response, % ( <i>n</i> )	34% (93)	43% (118)	44% (125)	58% (167)
Observed difference	-	9%	-	14%
Odds ratio (95% CI)	-	1.5 (1.1, 2.1)	-	1.8 (1.3, 2.6)

Analysis results of adult studies.

## Case Study 11: Pediatric Extrapolation with Robust MAP

- The final step of the process was to use a mixture prior approach to reweight the results to ensure that the adult data did not overwhelm the pediatric data in the analysis.
- This approach was used to vary the amount of information borrowing between no borrowing (represented by a weight of zero or 0) and full borrowing (represented by a weight of 1) where the adult and pediatric data are essentially pooled together, with every patient counted equally.
- The Bayesian analysis was performed for the entire range of weights, from no borrowing to full borrowing, to allow a complete view of the spectrum of outcomes.
- Point estimates (posterior means) and uncertainty intervals (95% credible intervals) were computed for each weight value.

## Case Study 11: Pediatric Extrapolation with Robust MAP



Posterior means and credible intervals for the pediatric belimumab study with varying prior weights.

## Case Study 11: Pediatric Extrapolation with Robust MAP

- Usually, it is best to prespecify the amount of borrowing.
- Although the mixture weight was not prespecified in this example, the clinical team did have a pre-existing belief that the effect of the treatment would be similar due to the PKPD of the disease.
- A prior weight of  $p_0 = 0.55$  resulted in a significant odds ratio (a posterior probability larger than 97.5%).
- The effective sample size was considered and was deemed acceptable by the review team.

## Bayesian Design

- In the Frequentist paradigm, asymptotics are typically used to control for the type I error rate at level  $\alpha$ .
- E.g., suppose the null hypothesis is  $H_0 : \theta \leq 0$ .
- For example,  $p$ -values ( $\Pr(D|\theta \leq 0)$ ) are shown to be asymptotically uniform, hence the usual cutoff  $\alpha = 0.05$ .
- The asymptotic distribution of these  $p$ -values is based on the **sampling distribution** of the test statistic under the null hypothesis  $H_0$ .
- It turns out that posterior probabilities,  $\Pr(\theta \leq 0|D)$  are also asymptotically uniform if the truth is  $\theta = 0$ .
- However, one of the advantages of the Bayesian paradigm is that inference is **exact**, so that we wish to avoid asymptotics wherever possible.

## Basics of Bayesian design

## Using Sampling and Fitting Priors

# Using Sampling and Fitting Priors

- Wang and Gelfand (2002) [22] proposed a general, simulation-based Bayesian SSD methodology based on analyst-specified “sampling” (i.e., design) and a “fitting” (i.e., analysis) priors.
- The **sampling prior** is a distribution over plausible values of the “truth.”
  - ▶ For example, if an analyst believes with 95% probability the response probability will be between 0.2 and 0.4, we may elicit

$$\pi_s(\theta) = \text{Beta}(\theta|23.64, 55.16).$$

- ▶ In general, the **sampling prior** will be informative.
- ▶ As a special case, to mimic frequentist sample size calculations, we may utilize a **point mass sampling prior**, e.g.,

$$\pi_s(\theta) = 1\{\theta = 0.3\}$$

# Using Sampling and Fitting Priors

- The **fitting prior** is the prior that will be used to analyze the **future** data set.

- ▶ The **fitting prior** may be informative or noninformative.
- ▶ E.g., if we wish to conduct Bayesian analysis but have no prior information, we may elicit a uniform prior, i.e.,

$$\pi_f(\theta) \propto 1\{0 \leq \theta \leq 1\}.$$

- ▶ If we have historical data, we may use a power prior as the fitting prior, i.e.,

$$\pi_f(\theta) \propto L(\theta|D_0)^{a_0} \pi_0(\theta).$$

## Using Sampling and Fitting Priors

- One generates a value of  $\theta$  from the **sampling prior**  $\pi_s(\theta)$ , and then generates a data set  $\mathbf{y}$  based on that  $\theta$ . This process produces sample data sets  $\mathbf{y}$  from their prior predictive distribution

$$p_s(\mathbf{y}) = \int f(\mathbf{y}|\theta) \pi_s(\theta) d\theta.$$

- The goal is to choose  $n$  to achieve expected behavior of the posterior

$$\pi_f(\theta|\mathbf{y}) \propto f(\mathbf{y}|\theta) \pi_f(\theta),$$

where  $\pi_f(\theta)$  denotes the **fitting prior**.

## Using Sampling and Fitting Priors

- Let  $T(\mathbf{y})$  be a function of  $\mathbf{y}$  which will be used to define the criterion that governs the appropriate value of  $n$ .
  - APVC: Take  $T(\mathbf{y}) = \text{Var}_f(\theta|\mathbf{y})$  and find  $n$  to ensure  $E_s[T(\mathbf{y})] \leq \epsilon$ .
  - ALC: Take  $T(\mathbf{y})$  to be the length of the  $100(1 - \alpha)\%$  posterior credible interval and find  $n$  to ensure  $E_s[T(\mathbf{y})] \leq w$ .
  - Assurance: Take  $T(\mathbf{y}) = 1\{P(\theta > \theta_0 | \mathbf{y}) > 1 - \alpha\}$  and find  $n$  to ensure  $E_s[T(\mathbf{y})] \geq 1 - \beta$ .
    - ★ Assurance is also called probability of success (POS), Bayesian expected power, or simply Bayesian power.

Bayesian type I error rate and power

## Bayesian power

- Suppose that  $y \sim \text{Ber}(\theta)$  and suppose that we wish to test

$$H_0 : \theta \leq \theta_0 \text{ versus } H_1 : \theta > \theta_0$$

- A frequentist may wish to determine sample size to have a high chance to reject  $H_0$  by the  $p$ -value by choosing some  $\theta_1$  as the “truth.”
- One approach to determine this sample size is via simulation. For a given  $n$ :
  - Generate  $y \sim \text{Binomial}(n, \theta_1)$
  - Check if the  $p$ -value  $< \alpha/2$  for testing  $H_0$  versus  $H_1$ .
  - Repeat many times, taking the proportion of rejected hypotheses until a desired power is achieved.

## Bayesian power

- An analogous approach can be conducted using sampling priors.
- In particular, we may set

$$\pi_s^{(H_1)}(\theta) = 1\{\theta = \theta_1\}.$$

- For a given  $n$ , the predictive distribution is thus given by

$$f_s^{(H_1)}(y) = \int \binom{n}{y} \theta^y (1-\theta)^{n-y} 1\{\theta = \theta_1\} d\theta = \binom{n}{y} \theta_1^y (1-\theta_1)^{n-y}.$$

## Bayesian power

- In practice, we do not know the true value of  $\theta$ , but we often have some prior information on what it may be (e.g., through previous studies).
- For example, if a previous study had  $y_0 = 8$  responders out of  $n_0 = 10$  participants, we may select  $\theta_1 = 0.8$ .
- However, this ignores uncertainty in the empirical estimate.
- Alternatively, we may elicit

$$\pi_s^{(H_1)}(\theta) \propto \theta^{8-1}(1-\theta)^{2-1},$$

which reflects our uncertainty.

- Hence, **Bayesian power** is a generalization of frequentist power.

## Bayesian power

- After eliciting a sampling prior, we must elicit a **fitting prior**.
- The **fitting prior** may be taken to be informative or non-informative.
- A common metric for the success criterion is the posterior probability exceeding some threshold (e.g.,  $1 - \alpha/2$ , i.e.,

$$\Pr(\theta > \theta_0 | \mathbf{y}) = \int_{\theta_0}^1 p(\theta | \mathbf{y}, \pi^{(f)}) \geq 1 - \alpha/2. \quad (13)$$

- This gives rise to the following algorithm to compute Bayesian power:
  - Obtain a sample of size  $n$  from the prior predictive distribution of  $\pi_s^{(H_1)}$ .
  - Using the generated data set and the **fitting prior**, compute the posterior probability in (13).
  - Repeat many times, computing the proportion of times the posterior probability exceeds  $1 - \alpha/2$ .

## Bayesian type I error rate

- In general, there is a tradeoff between the power gains associated with using informative priors and type I error inflation.
- It is important to note that, in finite samples, frequentist approaches have a type I error rate that may be above or below the nominal level  $\alpha$ .
- Similar to computing power, we may use the simulation-based approaches discussed so far to compute the Bayesian (i.e., finite sample) type I error rate.

## Bayesian type I error rate

- Consider the following sampling priors, which reflect a null and alternative hypothesis.

$$\begin{aligned}\pi_s^{(H_0)}(\theta) &= 1\{\theta = \theta_0\}, \\ \pi_s^{(H_1)}(\theta) &= 1\{\theta = \theta_1\}.\end{aligned}$$

- Sampling from the prior predictive distributions ( $f_s^{(H_0)}$  and  $f_s^{(H_1)}$ ) of these sampling priors yields data generated from  $\theta_0$  and  $\theta_1$  as the truth, respectively.
- Given a sample size  $n$ , we may compute the Bayesian type I error rate and power based on threshold  $\gamma$  as

$$\alpha_n = E_{f_s^{(H_0)}(\mathbf{y})} [1\{P(\theta > \theta_0 | \mathbf{y}) \geq \gamma\}] = \text{Bayes type I error}$$

$$\beta_n = E_{f_s^{(H_1)}(\mathbf{y})} [1\{P(\theta > \theta_0 | \mathbf{y}) \geq \gamma\}] = \text{Bayes power}$$

- If we choose  $\gamma = 1 - \alpha/2$ ,  $\alpha_n \rightarrow \alpha/2$  as  $n \rightarrow \infty$ .

# Bayesian Power and Type I Error Calculation: Bernoulli Proportion Example

- Suppose we wish to estimate power and type I error for a single-arm trial.
- We assume  $y_i \sim \text{Ber}(\theta)$ .
- We wish to test  $H_0 : \theta \geq \theta_0$  versus  $H_1 : \theta < \theta_0$ , where  $\theta_0 = 0.5$ .
- We consider the following sampling and fitting priors

Hypothesis	Sampling	Fitting
$H_1$	$\theta \sim \text{Beta}(2, 8)$	$U(0, 1)$
$H_1$	$\theta \sim \text{Beta}(2, 8)$	$\text{Beta}(2, 8)$
$H_1$	$\theta = 0.2$	$U(0, 1)$
$H_1$	$\theta = 0.2$	$\text{Beta}(2, 8)$
$H_0$	$\theta = 0.5$	$\text{Beta}(2, 8)$
$H_0$	$\theta = 0.5$	$U(0, 1)$

# Bayesian Power and Type I Error Calculation: Bernoulli Proportion Example

- The results of the simulation are located in [Examples/bayesPowerType1Error.html](#).
- We see that a non-degenerate sampling prior yields a larger sample size for the same level of power compared to  $\pi_s(\theta) = 1\{\theta = 0.5\}$ .
- Using an informative fitting prior yields higher power, at the cost of inflated type I error rates.
- In general, any informative prior will have an inflated type I error rate.
- The Bayes power never reaches 1. This is because the sampling prior does not preclude  $\theta = \theta_0 = 0.5$ .
- An alternative is to use Bayesian **conditional** expected power via the sampling prior

$$\tilde{\pi}_s^{(H_1)}(\theta) \propto \pi_s^{(H_1)}(\theta)1\{\theta < \theta_0\}.$$

## Example: Belimumab Adult SLE Program

- The BLISS-76 study was a phase III, multicenter, randomized, placebo-controlled trial in serologically active systemic lupus erythematosus (SLE) adult patients [23].
- Patients were randomized in 1:1:1 ratio to receive 1 mg/kg belimumab, 10 mg/kg belimumab, or placebo.
- The target sample size of 810 patients (270/group) was chosen to provide  $\geq 90\%$  power at the 5% significance level to detect a  $\geq 14\%$  absolute improvement in SRI response rate for belimumab 10 mg/kg relative to placebo.

## Example: Belimumab Adult SLE Program

- The population studied in BLISS-76 was selected based on a subgroup analysis ( $n = 321$ ) of phase II data [24].

	Placebo (n=86)	Belimumab				<i>P</i> value <sup>b</sup>	
		1.0 mg/kg (n=78)	4.0 mg/kg (n=79)	10.0 mg/kg (n=78)	All Active (n=235)		
<b>SRI response rate,<sup>c</sup> %</b>		29.1	48.7	43.0	46.2	46.0	0.006
4-point reduction in SELENA-SLEDAI		39.5	52.6	48.1	47.4	49.4	0.117
No worsening by BILAG <sup>d</sup>		81.4	88.5	94.9	91.0	91.5	0.015
No worsening by PGA <sup>e</sup>		76.7	89.7	88.6	92.3	90.2	0.003
<b>Modified SRI<sup>f</sup></b>	(n=74)	(n=70)	(n=67)	(n=68)	(n=205)		
Baseline SELENA-SLEDAI score, mean		8.2	8.1	8.2	8.1	8.2	
<b>Modified SRI response rate,<sup>f</sup> %</b>		32.4	48.6	46.3	47.1	47.3	0.025

## Example: Belimumab Adult SLE Program

- Model:  $y_g | \pi_g \sim \text{Bin}(\pi_g, n_g)$  for  $g = 1$  (belimumab) and  $g = 0$  (pbo).
- Prior:  $\pi_g \sim \text{beta}(0.5, 0.5)$  ([Jeffreys' prior](#) ).
- Posterior:  $\pi_g | y_g \sim \text{beta}(y_g + 0.5, (n_g - y_g) + 0.5)$ .
- It is straightforward to directly sample from the posterior distribution for  $(\pi_0, \pi_1)$  and then compute the posterior distribution  $\theta = \pi_1 - \pi_0$  from those samples.

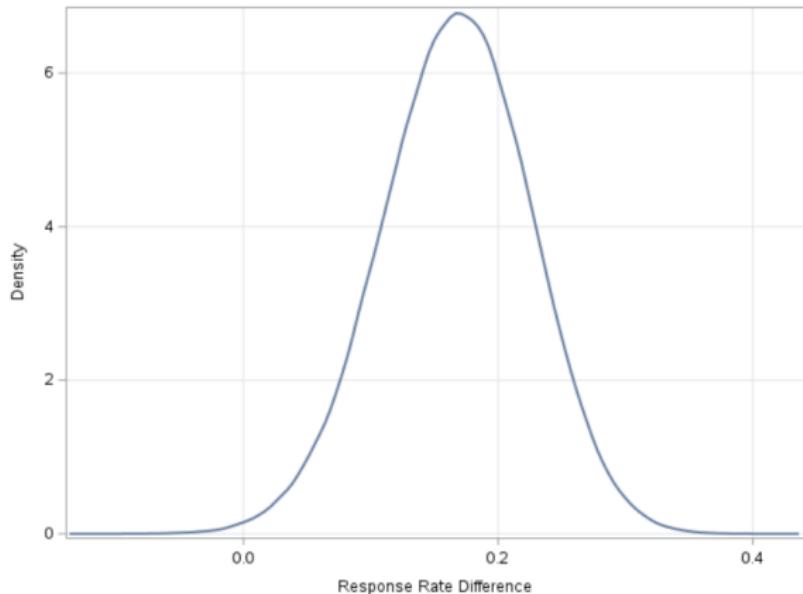
## Example: Belimumab Adult SLE Program

```
** input Phase II subgroup data;
data subGroup;
  infile datalines;
  input y0 n0 y1 n1;
  datalines;
  25 86 108 235
;
run;

** directly sample from posterior;
data postSamples;
  set subGroup;
  call streaminit(354352);

  do sample = 1 to 1000000;
    pi0    = rand('beta',y0+0.5,n0-y0+0.5);
    pi1    = rand('beta',y1+0.5,n1-y1+0.5);
    theta = pi1 - pi0;
    output;
  end;
run;
```

## Example: Belimumab Adult SLE Program



# Example: Belimumab Adult SLE Program

```
** estimate Bayesian power (asymptotic approximation);
data preposteriorAnalysis;
set postSamples(where=(theta>0));
call streaminit(12341);

n1      = 270;
y1      = rand('binomial',pi1,n1);
pi1Hat = (y1+0.5)/(n1+1.0);

n0      = 270;
y0      = rand('binomial',pi0,n0);
pi0Hat = (y0+0.5)/(n0+1.0);

muPost  = pi1Hat - pi0Hat;
sdPost   = sqrt(pi1Hat*(1-pi1Hat)/(n1+1.0) + pi0Hat*(1-pi0Hat)/(n0+1.0));
pstProb  = sdf('normal',0,muPost,sdPost);
rejNull = (pstProb>0.975);
run;

title "Operating Characteristics";
proc means data = preposteriorAnalysis mean;
var pi1Hat pi0Hat muPost sdPost pstProb rejNull;
run;
```

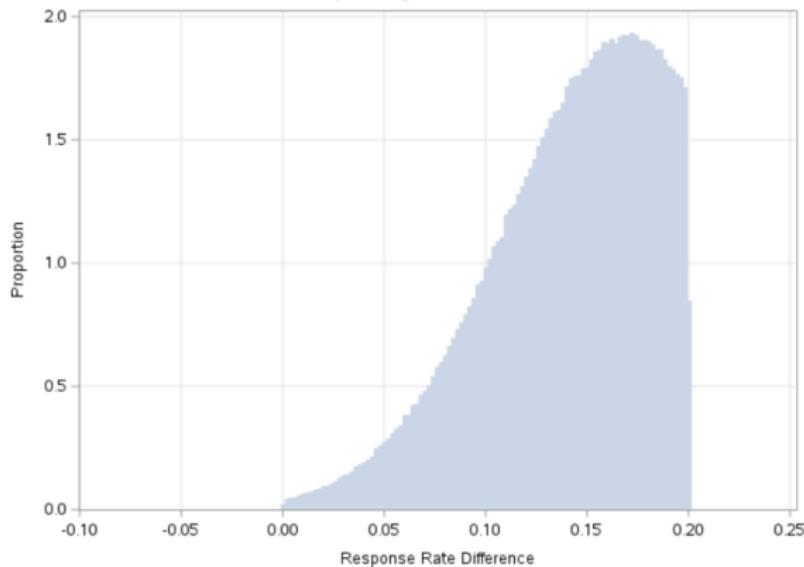
Operating Characteristics

Variable	Mean
pi1Hat	0.4600144
pi0Hat	0.2934643
muPost	0.1665502
sdPost	0.0407705
pstProb	0.9784396
rejNull	0.6834878

## Example: Belimumab Adult SLE Program

- Suppose we were concerned that the subgroup analysis provides an overly optimistic picture of the treatment effect.
- We consider the truncated alternative prior proposed by Psioda and Ibrahim [8].
- For this case, we restrict  $\theta \in (0.00, 0.20)$  to discount the overly optimistic range of effects.

## Example: Belimumab Adult SLE Program



- The Bayes power in this case drops to 83.6%.
- One may also wish to rule out very small values of the effect.

## Example: Belimumab Adult SLE Program

- Suppose now we want to find the required number of patients per group so that the study has 90% Bayesian power with respect to the truncated alternative sampling prior.

## Example: Belimumab Adult SLE Program

```
data postSamples2;
set subGroup;
call streaminit(354352);
do sample = 1 to 1000000;
    theta = 1;
    do until(0.00<theta<0.20);
        pi0    = rand('beta',y0+0.5,n0-y0+0.5);
        pi1    = rand('beta',y1+0.5,n1-y1+0.5);
        theta = pi1 - pi0;
    end;
    output;
end;
run;
```

## Example: Belimumab Adult SLE Program

```
%macro sample_size(nLower=10,nUpper=10,nBy=10,obs=100000);
%do n = &nLower. %to &nUpper. %by &nBy.;
  data preposteriorAnalysis2;
    set postSamples2(obs=&obs.);
    call streaminit(&n.);

    ** simulate data;
    n      = &n.;
    y1    = rand('binomial',pi1,n);
    y0    = rand('binomial',pi0,n);

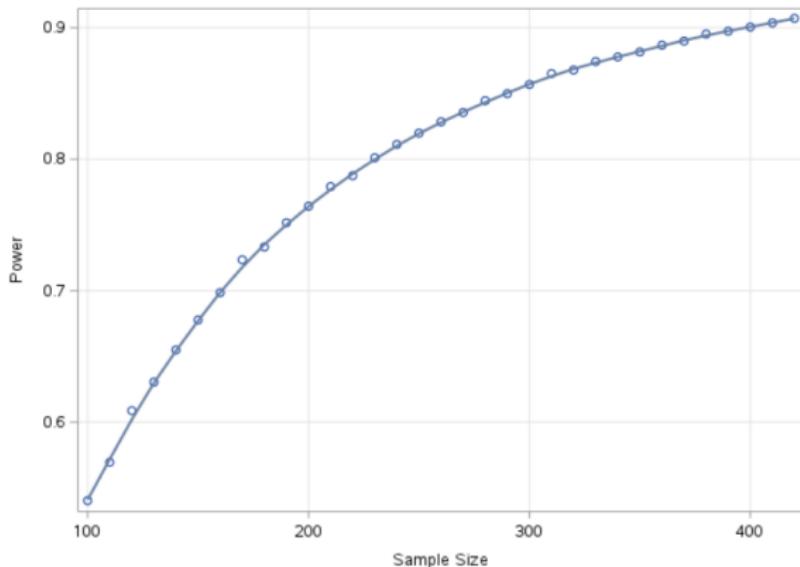
    ** approximate posterior;
    pi1Hat = (y1+0.5)/(n+1.0);
    pi0Hat = (y0+0.5)/(n+1.0);
    muPost = pi1Hat - pi0Hat;
    sdPost = sqrt(pi1Hat*(1-pi1Hat)/(n+1.0) + pi0Hat*(1-pi0Hat)/(n+1.0));

    ** calculate rejection rule;
    rejNull = (sdf('normal',0,muPost,sdPost)>0.975);
  run;

  proc means data = preposteriorAnalysis2 noprint;
  by n;
  var rejNull;
  output out = OperatingChar mean = bayesPower;
  run;
  proc append data = OperatingChar base = OperatingCharRange force; run; quit;
%end;
%mend;
```

## Example: Belimumab Adult SLE Program

```
%sample_size(nLower=100,nUpper=420,nBy=10);
title "Bayesian Power for Truncated Alternative Prior";
proc sgplot data = OperatingCharRange noautolegend;
    loess x=n y=bayesPower ;
    yaxis label = 'Power' grid;
    xaxis label = 'Sample Size' grid ;
run;
quit;
```



## Example: Pediatric Trial Design

- The original PLUTO trial used 5:1 randomization for the first 24 participants and 1:1 randomization thereafter for a target total sample size of  $n = 100$  ( $n_1 = 58$  treated,  $n_0 = 42$  placebo).
- For this example, we consider sampling priors

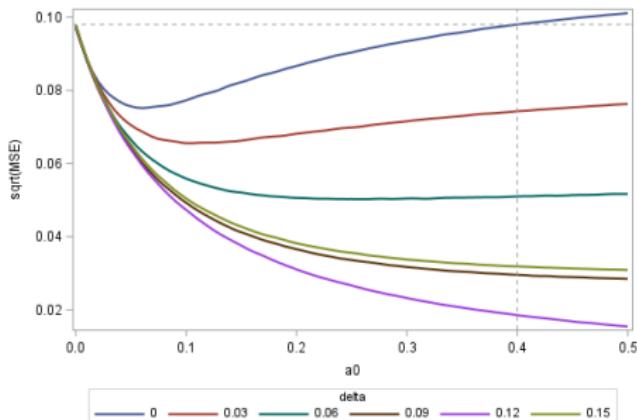
$$\pi_s(\delta, \theta_{0,c}) = \pi_s(\theta_{0,c}) \times 1[\theta_{1,c} = \theta_{0,c} + \delta]$$

for select values of  $\delta \in \{0.00, 0.03, 0.06, 0.09, 0.12, 0.15\}$ .

- We assume  $\pi_s(\theta_{0,c}) \sim \text{Beta}(218, 344)$  based on the pooled data from the adult trials (even though information is not borrowed on  $\theta_{0,c}$ ).
- To allow for computationally efficient design simulations, we exploit the asymptotic approximation to the partial-borrowing power prior as described in Psioda and Ibrahim [8].

## Example: Pediatric Trial Design

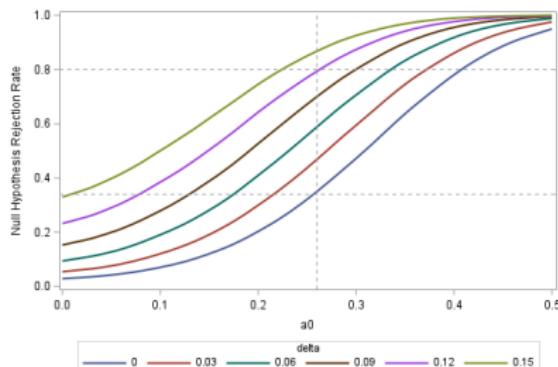
- See the [Examples/PLUTO\\_powerPriorDesign.sas](#) program.
- First, we examine the mean squared error (MSE) of the posterior mean estimator for  $\delta$ .



- One can see that taking  $a_0 < 0.40$  results in decreased MSE (relative to the no borrowing case) even when there are systematic differences in  $\delta$  across populations (e.g., no treatment effect in pediatric setting).

## Example: Pediatric Trial Design

- Next, we examine testing  $H_0 : \delta \leq 0$  versus  $H_0 : \delta > 0$ .



- For 80% power to detect an effect when  $\delta = 0.12$ , we need  $a_0 = 0.26$ .
- Note that for  $a_0 < 0.40$ , there is essentially zero probability of observing data that produces substantial evidence of benefit without there being *some* evidence in the pediatric data itself.
- Thus, specification of  $a_0 = 0.26$  would seem reasonable.

## Example: Pediatric Trial Design

- After identifying  $a_0 = 0.26$  using asymptotic approximations, we verify the design's properties using exact MCMC.
  - See the [Examples/PLUTO\\_powerPriorDesignMcmc.sas](#) program.
  - Note it can take several hours to complete the full set of simulation studies to reproduce the graphics with reasonable precision.
- We also compute the assurance based on  $\pi_s(\theta_{0,c}) \sim \text{Beta}(218, 344)$  and  $\pi_s(\theta_{1,c}) \sim \text{Beta}(285, 275)$ .
  - See the [Examples/PLUTO\\_powerPriorDesignMcmcAssurance.sas](#) program.
  - Note it can take several hours to complete 2500 simulation studies (using 20,000 MCMC samples per analysis).

# Data-Driven Priors and Clinical Trial Design

- Any of the informative priors discussed can be embedded in a trial's design as the fitting prior. The power prior is simply a popular choice.
- For example, one can evaluate the operating characteristics of a design that uses a commensurate prior as the fitting prior to determine if the design's operating characteristics are acceptable (for a set of sampling priors).
- Often the prior (or some other aspect of the design) may need be tuned to ensure operating characteristics are acceptable to regulatory and non-regulatory stakeholders.

## Recap of Simulation-Based SSD Procedure

- The previous examples illustrate the fundamental steps for Bayesian sample size determination.
- Set  $M = \text{number of simulated data sets}$ .
- For  $m = 1, \dots, M$  and a possible choice for  $n$ ,
  - ① Sample  $\theta_m$  from  $\pi_s(\theta)$ .
  - ② Simulate  $\mathbf{y}_m$  from  $p(\mathbf{y}_m|\theta_m)$ .
  - ③ Compute  $T(\mathbf{y}_m)$  as some function of  $p(\theta|\mathbf{y}_m)$ , possibly using MCMC, e.g.,
- If the average value of the  $T(\mathbf{y}_m)$  over  $m = 1, \dots, M$  satisfies the requirement then use  $n$  as the sample size, else increment  $n$  and repeat!

Other topics in Bayesian SSD

# Bayesian SSD

- For Bayesian sample size determination in such settings, commonly used criteria include the following:
  - ▶ Average Coverage Criterion (ACC)
  - ▶ Average Length Criterion (ALC)
  - ▶ Worst Outcomes Criterion (WOC)

## Bayesian SSD: Average Coverage Criterion

- For a fixed posterior interval of width  $w$ , we determine the sample size by finding the smallest  $n$  such that the following equation is satisfied.

$$\int \left( \int_{a(\mathbf{y}, n)}^{a(\mathbf{y}, n) + w} p(\theta | \mathbf{y}) d\theta \right) p(\mathbf{y}) d\mathbf{y} \geq 1 - \alpha$$

- Here,
  - $\Pr [\theta \in (a(\mathbf{y}, n), a(\mathbf{y}, n) + w) | \mathbf{y}] = \int_{a(\mathbf{y}, n)}^{a(\mathbf{y}, n) + w} p(\theta | \mathbf{y}) d\theta$
  - $p(\mathbf{y}) = \int p(\mathbf{y} | \theta) \pi(\theta) d\theta$  is the prior predictive distribution of  $\mathbf{y}$ .
- The ACC ensures that the mean coverage of posterior credible intervals of width  $w$ , weighted by  $p(\mathbf{y})$ , is at least  $1 - \alpha$ .
- The quantity  $a(\mathbf{y}, n)$  can be chosen to yield credible or HPD intervals.

## Bayesian SSD: Average Length Criterion

- For a fixed posterior credible interval of coverage  $1 - \alpha$ , we can determine the sample size by finding the smallest  $n$  such that

$$\int w(\mathbf{y}, n) p(\mathbf{y}) d\mathbf{y} \leq w$$

where  $w(\mathbf{y}, n)$  is the width of the  $100(1 - \alpha)\%$  posterior credible interval for data  $\mathbf{y}$ , determined by solving

$$\int_{a(\mathbf{y}, n)}^{a(\mathbf{y}, n) + w(\mathbf{y}, n)} p(\theta | \mathbf{y}) d\theta = 1 - \alpha$$

for  $w(\mathbf{y}, n)$  for each value of  $\mathbf{y} \in \mathcal{Y}$ .

- The ALC ensures that the mean length of the  $100(1 - \alpha)\%$  posterior credible intervals weighted by  $p(\mathbf{y})$  is at most  $w$ .

## Bayesian SSD: Worst Outcomes Criterion

- Cautious investigators may not be satisfied with the *average* assurances provided by the ACC and the ALC criteria.
- Therefore, a conservative sample size can also be determined by

$$\inf_{\mathbf{y} \in \mathcal{Y}^*} \left( \int_{a(\mathbf{y}, n)}^{a(\mathbf{y}, n) + w(\mathbf{y}, n)} p(\theta | \mathbf{y}) \, d\theta \right) \geq 1 - \alpha,$$

where  $\mathcal{Y}^* \subset \mathcal{Y}$  is a suitably chosen subset of the sample space.

- For example, the WOC ensures that if  $\mathcal{Y}^*$  consists of the most likely 95% of the possible  $\mathbf{y} \in \mathcal{Y}$ , then there is a 95% assurance that the length of the  $100(1 - \alpha)\%$  posterior credible interval will be at most  $w$ .

## Bayesian SSD: Single Normal Mean

Let us first consider a one-sample problem and a normal model. Here we can derive closed form expressions.

- Suppose  $y_1, \dots, y_n$  are *i.i.d.*  $N(\mu, \sigma^2)$ .
- Let  $\tau = 1/\sigma^2$  represent the precision parameter.
- Assume the usual conjugate priors for  $(\mu, \tau)$ .

$$\mu | \tau \sim N\left(\mu_0, \frac{1}{n_0 \tau}\right) \quad \text{and} \quad \tau \sim \text{gamma}(\alpha_0, \lambda_0)$$

- We use the rate parameterization so that  $E(\tau) = \frac{\alpha_0}{\lambda_0}$  and  $\text{Var}(\tau) = \frac{\alpha_0}{\lambda_0^2}$ .

## Bayesian SSD: Single Normal Mean

- Case 1:  $\tau$  known: In the case that  $\tau$  is known, we have

$$\mu | \mathbf{y} \sim N(\mu_n, \tau_n^{-1})$$

where

$$\mu_n = \frac{n_0 \mu_0 + n \bar{y}}{n_0 + n} \quad \text{and} \quad \tau_n = (n + n_0) \tau.$$

## Bayesian SSD for a Normal Mean

- Since the posterior precision of  $\mu$  depends only on  $n$  and does not vary with the observed data vector  $\mathbf{y}$ , all three criteria (ACC, ALC, WOC) lead to the same solution, which is

$$n \geq \frac{4z_{1-\alpha/2}^2}{\tau w^2} - n_0, \quad (14)$$

where  $w$  is the desired width of the posterior interval for  $\mu$ .

- If a uniform improper prior for  $\mu$  is used (i.e.,  $n_0 = 0$ ), then inequality (14) reduces to the frequentist formula for the sample size in (??).

# Bayesian SSD for a Normal Mean

- Case 2:  $\tau$  unknown: If  $\tau$  is unknown, then

$$\mu | \mathbf{y} \sim t \left( n + n_0, \mu_n, \frac{2\beta_n}{(n + 2\alpha_0)(n + n_0)} \right),$$

i.e.,

$$\mu | \mathbf{y} \stackrel{d}{=} \mu_n + \sqrt{\frac{2\beta_n}{(n + 2\alpha_0)(n + n_0)}} T_{n+n_0},$$

where

$$\beta_n = \lambda_0 + \frac{n}{2}s^2 + \frac{nn_0}{2(n + n_0)}(\bar{y} - \mu_0)^2,$$

$ns^2 = \sum_{i=1}^n (y_i - \bar{y})^2$ , and  $T_{n+n_0}$  is a Student's  $t$  random variable with  $n + n_0$  degrees of freedom

- Since the posterior precision of  $\mu$  varies depends on the observed data  $\mathbf{y}$ , different criteria will lead to different sample sizes.

## Bayesian SSD for a Normal Mean (ACC)

- For the ACC, [25] shows the formula for the sample size is given by

$$n \geq \left( \frac{4\lambda_0}{\alpha_0 w^2} \right) t_{(2\alpha_0, 1-\alpha/2)}^2 - n_0. \quad (15)$$

- Since  $\alpha_0/\lambda_0$  is the prior mean for  $\tau$ , the ACC sample size for unknown  $\tau$  is similar to that for known  $\tau$ , in that we only need to substitute the prior mean precision for  $\tau$  in inequality (14) and exchange the normal quantile  $z$  with a quantile from a  $t_{2\alpha_0}$  distribution.
- Since the degrees of freedom of the  $t$  distribution in (15) do not increase with the sample size, the ACC can lead to sample sizes that are substantially different from those in (14) or (??).

## Bayesian SSD for a Normal Mean

- For the ALC, it can be shown that the required sample size satisfies

$$2t_{(n+2\alpha_0, 1-\alpha/2)} \left( \frac{2\lambda_0}{(n + 2\alpha_0)(n + n_0)} \right)^{1/2} \frac{\Gamma\left(\frac{n+2\alpha_0}{2}\right) \Gamma\left(\frac{2\alpha_0-1}{2}\right)}{\Gamma\left(\frac{n+2\alpha_0-1}{2}\right) \Gamma(\alpha_0)} \leq w \quad (16)$$

- Although it is not feasible to solve the inequality in (16) explicitly for  $n$ , it is straightforward to calculate given  $\alpha_0, \lambda_0, n_0, \alpha$ , and  $w$ .
- For example, since  $n$  is the only parameter that varies, one can conduct a grid search to solve the inequality.

## Example: Bayesian SSD for a Normal Mean

Prior Hyperparameters			Design Inputs		Bayesian SSD Criterion							
alpha0	lambda0	n0	Length	alpha	ACC	ALC	ALC/ACC	WOC	WOC/ACC	Freq.	Freq./ACC	
2	2	10	0.2	0.01	2110	1035	0.49	8924	4.23	664	0.31	
				0.05	761	595	0.78	2152	2.83	385	0.51	
				0.10	445	416	0.93	1007	2.26	271	0.61	
			0.5	0.01	330	160	0.48	1421	4.31	107	0.32	
				0.05	114	88	0.77	336	2.95	62	0.54	
				0.10	63	59	0.94	153	2.43	44	0.70	
100	100	10	0.2	0.01	667	661	0.99	822	1.23	664	1.00	
				0.05	379	378	1.00	436	1.15	385	1.02	
				0.10	264	263	1.00	292	1.11	271	1.03	
			0.5	0.01	99	98	0.99	116	1.17	107	1.08	
				0.05	53	53	1.00	58	1.09	62	1.17	
				0.10	34	34	1.00	36	1.06	44	1.29	

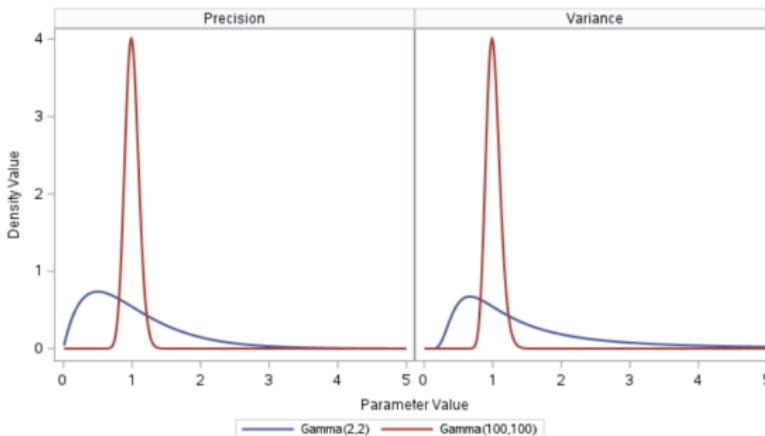
Note: the frequentist sample size satisfies  $n \geq 4z_{1-\alpha/2}^2 / \tau l^2$ , (i.e., equation (14) with  $n_0 = 0$ ).

## Example: Bayesian SSD for a Normal Mean

- Rows 1-6 of the table above show that the Bayesian approach can provide larger sample sizes than the frequentist approach, even though prior information is incorporated in the final inferences.
- With a large amount of prior information on both  $(\mu, \tau)$ , the Bayesian approach leads to smaller sample sizes than the frequentist approach. (see rows for  $\alpha_0 = \lambda_0 = 100$ ).
- With an informative prior on  $\tau$ , but not on  $\mu$ , similar sample sizes are provided by all criteria, with the WOC criterion somewhat higher than the rest.

## Example: Bayesian SSD for a Normal Mean

- To appreciate the impact of having imprecise information about the unknown variance, consider its induced prior distribution.



- The induced prior for  $\sigma^2$  based on  $\tau \sim \text{Gamma}(2, 2)$  satisfies:
  - $P(\sigma^2 > 1) = 0.594 \rightarrow n_{\text{frequentist}}(\sigma^2 = 1) = 385$
  - $P(\sigma^2 > 4) = 0.090 \rightarrow n_{\text{frequentist}}(\sigma^2 = 4) = 1537$

## Summary

- The ACC, ALC, and WOC criteria provide an attractive alternative to frequentist sample size determination.
- When a non-informative prior is elicited for  $\mu$ , the Bayesian methods will typically yield a larger sample size.
- As the prior variance for  $\tau \rightarrow 0$ , the three criteria converge to the frequentist method for sample size.
  - ▶ Said differently, the frequentist methods can be considered as limiting cases of the Bayesian methods with variance known and fixed.
- A primary weakness of these methods is that they are only useful for conjugate priors yielding tractable posterior densities.

## Bayesian SSD for a Normal Mean

- For the WOC, it can be shown that  $n$  satisfies

$$\frac{w^2(n + 2\alpha_0)(n + n_0)}{8\lambda_0 \left(1 + \frac{n}{2\alpha_0} F_{(n, 2\alpha_0, 1-\alpha)}\right)} \geq t_{(n+2\alpha_0, 1-\alpha/2)}^2 \quad (17)$$

where  $F(n, 2\alpha_0, 1 - \alpha)$  denotes the  $100 \times (1 - \alpha)$  percentile of the  $F$  distribution with  $(n, 2\alpha_0)$  degrees of freedom.

- The smallest  $n$  satisfying (16) or (17) can be found by an optimization routine, e.g., a grid search.

## Example: Bayesian SSD for a Normal Mean

- ACC can be calculated directly.
- ALC and WOC are computed easily via grid search, solving for sample size  $n$  in equations (16) or (17).

## Selected References

## Selected References I

- [1] Ming-Hui Chen, Shan Qi-Man, and Joseph G Ibrahim. *Monte Carlo Methods in Bayesian Computation*. Springer-Verlag New York, 1 edition, 2000.
- [2] Ethan M Alt, Matthew A Psioda, and Joseph G Ibrahim. Bayesian multivariate probability of success using historical data with type i error rate control. *Biostatistics*, 24(1):17–31, 2023.
- [3] Persi Diaconis and Donald Ylvisaker. Conjugate priors for exponential families. *The Annals of statistics*, pages 269–281, 1979.
- [4] Mayetri Gupta and Joseph G Ibrahim. An information matrix prior for bayesian analysis in generalized linear models with high dimensional data. *Statistica Sinica*, 19(4):1641, 2009.
- [5] Gene Pennello and Laura Thompson. Experience with reviewing bayesian medical device trials. *Journal of Biopharmaceutical Statistics*, 18(1):81–115, 2007.

## Selected References II

- [6] Joseph G. Ibrahim and Ming Hui Chen. Power prior distributions for regression models. *Statistical Science*, 15:46–60, 2000.
- [7] Joseph G. Ibrahim, Ming Hui Chen, Yeongjin Gwon, and Fang Chen. The power prior: Theory and applications. *Statistics in Medicine*, 34:3724–3749, 12 2015.
- [8] Matthew A. Psioda and Joseph G. Ibrahim. Bayesian clinical trial design using historical data that inform the treatment effect. *Biostatistics*, 20:400–415, 2019.
- [9] Yuyan Duan, Keying Ye, and Eric P. Smith. Evaluating water quality using power priors to incorporate historical information. *Environmetrics*, 17:95–106, 2006.
- [10] Luiz Max Carvalho and Joseph G Ibrahim. On the normalized power prior. *Statistics in Medicine*, 40:5251–5275, 2021.

## Selected References III

- [11] Andrew Gelman. Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper). *Bayesian analysis*, 1(3):515–534, 2006.
- [12] Ming-Hui Chen and Joseph G Ibrahim. The relationship between the power prior and hierarchical models. *Bayesian Analysis*, 1(3):551–574, 2006.
- [13] Brian P. Hobbs, Daniel J. Sargent, and Bradley P. Carlin. Commensurate priors for incorporating historical information in clinical trials using general and generalized linear models. *Bayesian Analysis*, 7:639–674, 2012.
- [14] Ethan M Alt, Brady Nifong, Xinxin Chen, Matthew A Psioda, and Joseph G Ibrahim. The scale transformed power prior for use with historical data from a different outcome model. *Statistics in Medicine*, 42(1):1–14, 2023.

## Selected References IV

- [15] Chenguang Wang, Heng Li, Wei-Chen Chen, Nelson Lu, Ram Tiwari, Yunling Xu, and Lilly Q Yue. Propensity score-integrated power prior approach for incorporating real-world evidence in single-arm clinical studies. *Journal of Biopharmaceutical Statistics*, 29(5):731–748, 2019.
- [16] Nelson Lu, Chenguang Wang, Wei-Chen Chen, Heng Li, Changhong Song, Ram Tiwari, Yunling Xu, and Lilly Q Yue. Propensity score-integrated power prior approach for augmenting the control arm of a randomized controlled trial by incorporating multiple external data sources. *Journal of biopharmaceutical statistics*, pages 1–12, 2021.
- [17] David J Spiegelhalter, Nicola G Best, Bradley P Carlin, and Angelika Van Der Linde. Bayesian measures of model complexity and fit. *Journal of the royal statistical society: Series b (statistical methodology)*, 64(4):583–639, 2002.

## Selected References V

- [18] Aki Vehtari, Andrew Gelman, and Jonah Gabry. Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and computing*, 27(5):1413–1432, 2017.
- [19] Ming-Hui Chen and Joseph G Ibrahim. Conjugate priors for generalized linear models. *Statistica Sinica*, pages 461–476, 2003.
- [20] Ethan M Alt, Matthew A Psioda, and Joseph G Ibrahim. A hierarchical prior for generalized linear models based on predictions for the mean response. *Biostatistics*, 23(4):1165–1181, 2022.
- [21] Hong Zhao, Brian P Hobbs, Haijun Ma, Qi Jiang, and Bradley P Carlin. Combining non-randomized and randomized data in clinical trials using commensurate priors. *Health Services and Outcomes Research Methodology*, 16:154–171, 2016.

## Selected References VI

- [22] Fei Wang and Alan E Gelfand. A Simulation-Based Approach to Bayesian Sample Size Determination for Performance under a Given Model and for Separating Models. *Statistical Science*, 17(2):193–208, 2002.
- [23] Richard Furie, Michelle Petri, Omid Zamani, Ricard Cervera, Daniel J Wallace, Dana Tegzová, Jorge Sanchez-Guerrero, Andreas Schwarting, Joan T Merrill, W Winn Chatham, William Stohl, Ellen M Ginzler, Douglas R Hough, Z John Zhong, William Freimuth, Ronald F van Vollenhoven, and BLISS-76 Study Group. A phase iii, randomized, placebo-controlled study of belimumab, a monoclonal antibody that inhibits b lymphocyte stimulator, in patients with systemic lupus erythematosus. *Arthritis and Rheumatism*, 63:3918–3930, 12 2011.

## Selected References VII

- [24] Richard A Furie, Michelle A Petri, Daniel J Wallace, Ellen M Ginzler, Joan T Merrill, William Stohl, W Winn Chatham, Vibeke Strand, Arthur Weinstein, Marc R Chevrier, Z John Zhong, and William W Freimuth. Novel evidence-based systemic lupus erythematosus responder index. *Arthritis Care and Research*, 61:1143–1151, 9 2009. <https://doi.org/10.1002/art.24698>.
- [25] CJ Adcock. A bayesian approach to calculating sample sizes. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 37(4-5):433–439, 1988.