Leveraging Historical Information: Methods and Applications

Ming-Hui Chen

Department of Statistics, University of Connecticut
E-Mail: ming-hui.chen@uconn.edu

2020 ASA Biopharmaceutical Section Regulatory-Industry Statistics Workshop (Virtual)

September 24, 2020

Outline

Historical Information

- Historical data are often available in clinical trials, genetics, health care, psychology, environmental health, engineering, economics, and business.
- In medical devices, historical data are often available from previous trials only from the control device.
- In pediatric rare cancer study, the data from adult patients may be available.
- In rare disease setting, an efficacious standard of care (S) is already on the market. Thus, the historical data are available from the treatment of S.

Leveraging Historical Information: Power Prior

- The first paper to discuss the formalization of the power prior as a general prior for various classes of regression models is Ibrahim and Chen (2000).
- Chen and Ibrahim (2006) establish the relationship between the power prior and hierarchical models.
- Ibrahim et al. (2015) give an A to Z exposition of the power prior and its applications to date.
- The power prior has emerged as a useful class of informative priors for a variety of situations in which historical data are available.
- References

♢ Ibrahim, J. G. and Chen, M.-H. (2000). Power prior distributions for regression models. *Statistical Science* **15**, 46-60.

♢ Chen, M.-H. and Ibrahim, J.G. (2006). The Relationship Between the Power Prior and Hierarchical Models. *Bayesian Analysis* **1**, 551-574.

♢ Ibrahim, J.G., Chen, M.-H., Gwon, Y., and Chen, F. (2015). The Power Prior: Theory and Applications. *Statistics in Medicine, 34*, 3724-3749.

The Basic Setting for the Power Prior

- Let the data from the current study be denoted by $D = (n, y, X)$, where n denotes the sample size, y denotes the $n \times 1$ response vector, and X denotes the $n \times p$ matrix of covariates.
- Denote the likelihood for the current study by $L(\boldsymbol{\theta}|D)$, where $\boldsymbol{\theta}$ is the vector of model parameters. Thus, $L(\boldsymbol{\theta}|D)$ can be a general likelihood function for an arbitrary regression model, such as a generalized linear model, random effects model, nonlinear model, or a survival model with right censored data.
- Denote the historical data by $D_0 = (n_0, y_0, X_0)$.
- Let $\pi_0(\boldsymbol{\theta})$ denote the prior distribution for $\boldsymbol{\theta}$ before the historical data D_0 is observed.
- $\pi_0(\boldsymbol{\theta})$ is typically taken to be improper.
- $\pi_0(\boldsymbol{\theta})$ is called the initial prior distribution for $\boldsymbol{\theta}$.

Basic Formulation of the Power Prior

- Given a_0, the power prior (Ibrahim and Chen, 2000) of $\boldsymbol{\theta}$ for the current study is defined as

$$\pi(\boldsymbol{\theta}|D_0, a_0) \propto L(\boldsymbol{\theta}|D_0)^{a_0}\pi_0(\boldsymbol{\theta}).$$

- a_0 is a scalar prior parameter that weights the historical data relative to the likelihood of the current study. It controls the influence of the historical data on $\pi(\boldsymbol{\theta}|D_0, a_0)$.
- a_0 can be interpreted as a discounting parameter, a precision parameter, and a parameter which reflects the heterogeneity (compatibility) between current and historical data.
- It is reasonable to restrict the range of a_0 to be between 0 and 1, and thus we take $0 \leq a_0 \leq 1$.
- a_0 controls the heaviness of the tails of the prior for $\boldsymbol{\theta}$. As a_0 becomes smaller, the tails of $\pi(\boldsymbol{\theta}|D_0, a_0)$ become heavier.

Example: Logistic Regression Model

- We simulated a data set consisting $n_0 = 200$ independent Bernoulli observations with success probability

$$p_{0i} = \frac{\exp\{-0.5 + 0.5x_{0i}\}}{1 + \exp\{-0.5 + 0.5x_{0i}\}} , \qquad i = 1, 2, \ldots, n_0 ,$$

where the x_{0i} are $i.i.d.$ normal random variables with mean 0 and standard deviation 0.5.

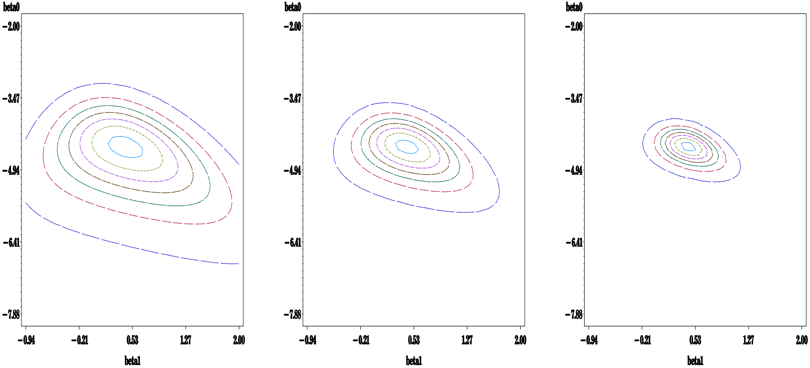
- Let $\boldsymbol{\beta} = (\beta_0, \beta_1)'$. Then, the likelihood function is given by

$$L(\boldsymbol{\beta}|D_0) = \prod_{i=1}^{n_0} \frac{\exp\{y_{0i}\boldsymbol{x}_{0i}'\boldsymbol{\beta}\}}{1 + \exp\{\boldsymbol{x}_{0i}'\boldsymbol{\beta}\}},$$

and the power prior with an improper uniform initial prior is thus given by

$$\pi(\boldsymbol{\beta}|D_0, a_0) \propto \prod_{i=1}^{n_0} \frac{\exp\{a_0 y_{0i}\boldsymbol{x}_{0i}'\boldsymbol{\beta}\}}{(1 + \exp\{\boldsymbol{x}_{0i}'\boldsymbol{\beta}\})^{a_0}}.$$

Figure: Contours of the Power Prior for $a_0 = 0.07, 0.17, 0.50$



- The centers of the power priors remain the same for different a_0 values.
- The tails of the power priors become heavier and the prior surfaces are getting flatter, as a_0 becomes smaller.

Normalized Power Priors

- Assuming that a_0 is random, the normalized power prior (Duan, Ye, and Smith, 2006; Neuenschwander et al., 2009) of $\boldsymbol{\theta}$ for the current study is defined as

$$\pi(\boldsymbol{\theta}, a_0|D_0) = \pi(\boldsymbol{\theta}|D_0, a_0)\pi(a_0) = \frac{L(\boldsymbol{\theta}|D_0)^{a_0}\pi_0(\boldsymbol{\theta})}{\int L(\boldsymbol{\theta}|D_0)^{a_0}\pi_0(\boldsymbol{\theta})d\boldsymbol{\theta}}\pi_0(a_0),$$

  where $\pi_0(\boldsymbol{\theta})$ is an initial prior and $\pi(a_0)$ is a marginal prior for a_0.
- For the normalized power prior, we must have

$$\int L(\boldsymbol{\theta}|D_0)^{a_0}\ \pi_0(\boldsymbol{\theta})d\boldsymbol{\theta} < \infty$$

  for $0 < a_0 \leq 1$.
- Ibrahim, Chen, Xia, and Liu (2012, Biometrics) propose the partial borrowing power prior.
- Hobbs et al. (2011, Biometrics; 2012, Bayesian Analysis) propose the hierarchical commensurate and power prior.

Outline

HPD Region

- For parameter $\boldsymbol{\theta}$ with the probability density function $f(\boldsymbol{\theta})$, the $100(1-\alpha)\%$ HPD region is the subset of the parameter space Θ such that
$$R(\alpha) = \{\boldsymbol{\theta} \in \Theta : f(\boldsymbol{\theta}) \geq f_\alpha\},$$
where f_α is the largest constant such that $P(\boldsymbol{\theta} \in R(\alpha)) \geq 1-\alpha$.
- Theorem: *For log-concave densities, the* $100(1-\alpha)\%$ *HPD region is a closed convex set.*

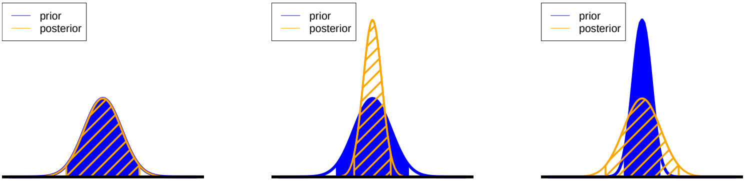
Notation

- Let $\pi(\boldsymbol{\theta}|\text{data}) \propto L(\boldsymbol{\theta}|D)\pi(\boldsymbol{\theta})$ denote the posterior distribution given the data D, where $L(\boldsymbol{\theta}|D)$ and $\pi(\boldsymbol{\theta})$ are the likelihood function and a prior distribution.
- The $100(1-\alpha)\%$ HPD region for $\boldsymbol{\theta}$ based on the prior distribution is then defined as
$$R_1(\alpha) = \left\{\boldsymbol{\theta} \in \Theta : \pi(\boldsymbol{\theta}) \geq \pi_\alpha^1\right\},$$
where π_α^1 is the largest constant such that $P(\boldsymbol{\theta} \in R_1(\alpha)) \geq 1-\alpha$.
- Similarly, the $100(1-\alpha)\%$ HPD region for $\boldsymbol{\theta}$ based on the posterior distribution is given by
$$R_2(\alpha) = \left\{\boldsymbol{\theta} \in \Theta : \pi(\boldsymbol{\theta}|D) \geq \pi_\alpha^2\right\},$$
where π_α^2 is the largest constant such that $P(\boldsymbol{\theta} \in R_2(\alpha)) \geq 1-\alpha$.

Information $\mathcal{I}$

- Our measure $\mathcal{I}$ is based on the comparison of $V(R_1(\alpha))$ and $V(R_2(\alpha))$, where $V(\cdot)$ represents the volume.
- Definition 1: Let $\phi = V(R_1(\alpha))/V(R_2(\alpha))$. The information $\mathcal{I}$ is defined as:

$$\mathcal{I} = \log \phi = \log \frac{V(R_1(\alpha))}{V(R_2(\alpha))}.$$



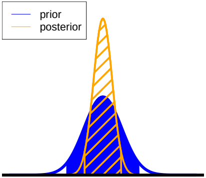
(a) No Information (b) Positive Information (c) Negative Information

Dissonance $\mathcal{D}$

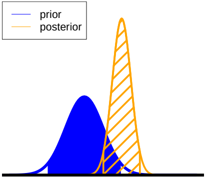
- Definition 2: For $R_1(\alpha)$ and $R_2(\alpha)$, let $R_{\text{min}}(\alpha)$ denote the region with smaller volume and $R_{\text{max}}(\alpha)$ denote the larger one. The dissonance $\mathcal{D}$ is measured as the fraction of the volume of the smaller HPD region that is not overlapping with the larger HPD region, i.e.,

$$\mathcal{D} = \frac{V(R_{\text{min}}(\alpha) \cap \overline{R_{\text{max}}(\alpha)})}{V(R_{\text{min}}(\alpha))}, \tag{1}$$

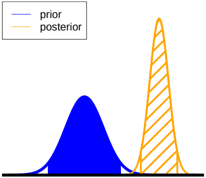
where $R_{\text{min}}(\alpha) \cap \overline{R_{\text{max}}(\alpha)} = \{\boldsymbol{\theta} \in \Theta : \boldsymbol{\theta} \in R_{\text{min}}(\alpha), \boldsymbol{\theta} \notin R_{\text{max}}(\alpha)\}$.



(d) No Dissonance



(e) Partial Dissonance



(f) Complete Dissonance

Comparing Two Data Sets or Two Posterior Distributions

- The two new measures $\mathcal{I}$ and $\mathcal{D}$ can be extended to compare two data sets or two posterior distributions given that the the parameter spaces are assumed to be the same.
- We can simply let $R_1(\alpha)$ and $R_2(\alpha)$ be the HPD regions computed under the two posterior distributions corresponding to two data sets using the same prior or corresponding to two prior distributions using the same data set.

Choice of α

- The values of $\mathcal{I}$ and $\mathcal{D}$ depend on the content level α.
- In practice, we need to choose α such that we consider using a $100(1-\alpha)\%$ HPD region to represent a set of plausible values.
- For our measure $\mathcal{I}$, one common choice is $\alpha = 0.05$, i.e., using the 95% HPD region.
- We can also compute $\mathcal{I}$ for different α values to get overall conclusion.
- Our measure $\mathcal{D}$ is more sensitive to the choice of α.
- Instead of fixing a value of α, we plot the curve of $\mathcal{D}$ versus α and summarize the extent of conflict using area under the curve (d−AUC), with smaller value suggesting that two data sets/two distributions are compatible and larger value indicating contradiction in the range of $[0, 1]$.

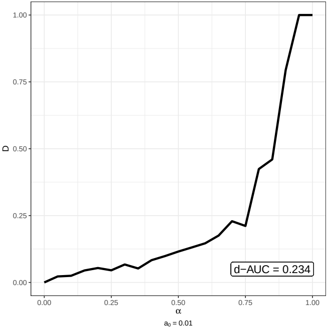
Application to Pediatric Cancer Data

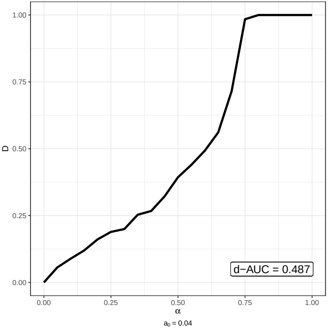
- The data are from Ye et al. (Pharmaceutical Statistics, 2020, DOI: 10.1002/pst.2039).
- Examine the effect of NDA22068 Nilotinib for pediatric patients.
- The outcome variable is the major molecular response (MMR: BCRABL/ABL $\leq 0.1\%$ IS).
- The data: $D_0 = (n_A = 282, y_A = 125)$ for adult patients and $D = (n_P = 25, y_P = 15)$ for pediatric patients.
- Assume that $y_A \sim B(n_A, p)$ and $y_P \sim B(n_P, p)$.
- Consider the power using D_A as the "historical data":

$$\begin{aligned} \pi(p) &\propto p^{-1}(1-p)^{-1} \\ \pi(p|D_0, D, a_0) &\propto \pi(p)[L(p|D_0)]^{a_0} L(p|D) \\ &\sim Beta(a_0 y_A + y_P, a_0(n_A - y_A) + (n_P - y_P)) \end{aligned}$$

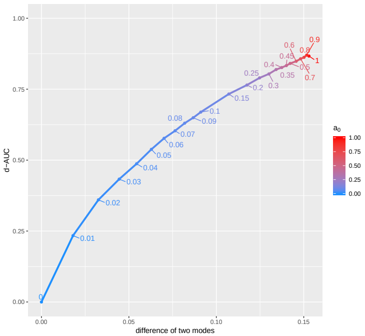
- The range of a_0 is between 0 and 1, with $a_0 = 0$ meaning that no incorporation of Adult data.
- We compare $\pi(p|D_0, D, a_0)$ to $\pi(p|D_0, D, a_0 = 0)$ via $\mathcal{D}$ and $\mathcal{I}$.

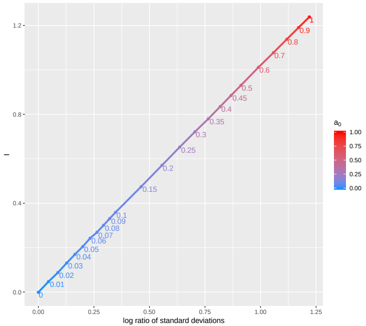
Plots of $\mathcal{D}$ versus α for 2 choices of a_0



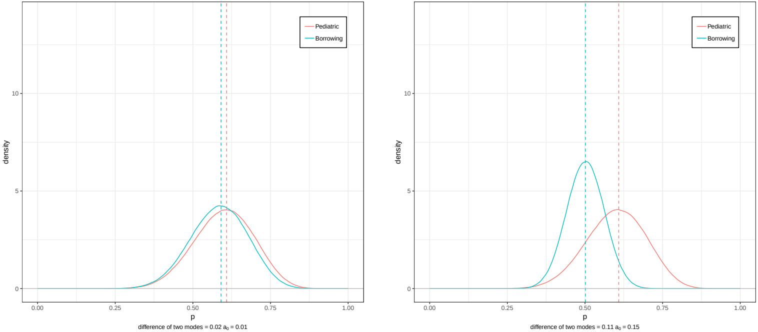


Plots of d-AUC and $\mathcal{I}$ over different choices of a_0





Posterior distributions of Pediatric Data Only versus Borrowing



- Wei, S., Chen, M.-H., Kuo. L., and Lewis, P.O. (2020+). Bayesian Information and Dissonance. under revision for *Bayesian Analysis*.

Outline

Notation, Historical Data and Hypotheses

- n_0 and n_c: the sample sizes of historical and current control arms; n_t ($> n_c$): the sample size of the test arm. We assume

$$y_{0i} \overset{i.i.d.}{\sim} N(\mu_c, \sigma_c^2), \quad y_{ci} \overset{i.i.d.}{\sim} N(\mu_c, \sigma_c^2), \text{ and } y_{ti} \overset{i.i.d.}{\sim} N(\mu_t, \sigma_t^2),$$

  where μ_c and σ_c are the mean and the standard deviation of the control arm, and μ_t and σ_t are the mean and the standard deviation of the test arm.

- Historical data

n_0	Mean	SD	Age
44	-0.18	3.38	4 to 8

- The means and standard deviations are in the unit of change in 6-month NSAA total score.
- Hypothesis of interest: H_0: $\delta = \mu_t - \mu_c \leq 0$ versus H_1: $\delta = \mu_t - \mu_c > 0$ for a superiority trial comparing test drug with the placebo. $\delta =$ the effect size.
- μ_c, σ_c^2, and σ_t^2 are nuisance parameters.

Posteriors with the Power Priors and Decision Rule

- $D_0 = (n_0, \bar{y}_0, S_0^2)$ and $D = (n_t, \bar{y}_t, S_t^2, n_c, \bar{y}_c, S_c^2)$, where $\bar{y}_0$, $\bar{y}_c$, and $\bar{y}_t$ are the sample means, and S_0^2, S_c^2, and S_t^2 are the sample variances for the historical data, the control and test arms, respectively.
- $\boldsymbol{\theta} = (\mu_c, \sigma_c^2, \mu_t, \sigma_t^2)'$.
- The posterior distribution with the power prior is given by

$$\pi(\boldsymbol{\theta}|D_0, D, a_0) \propto (\sigma_t^2)^{-\frac{n_t}{2}} \exp\Big\{ -\frac{1}{2\sigma_t^2}[n_t(\bar{y}_t - \mu_t)^2 + (n_t - 1)S_t^2]\Big\}$$
$$\times (\sigma_c^2)^{-\frac{n_c}{2}} \exp\Big\{ -\frac{1}{2\sigma_c^2}[n_c(\bar{y}_c - \mu_c)^2 + (n_c - 1)S_c^2]\Big\}$$
$$\times \Big((\sigma_c^2)^{-\frac{1}{2}} \exp\{-\frac{n_0(\bar{y}_0 - \mu_c)^2}{2\sigma_c^2}\}(S_0^2)^{\frac{n_0-3}{2}}(\sigma_c^2)^{-\frac{n_0-1}{2}} \exp\Big[-\frac{(n_0-1)S_0^2}{2\sigma_c^2}\Big]\Big)^{a_0}\pi_0(\boldsymbol{\theta}),$$

  where $\pi_0(\boldsymbol{\theta})$ is an initial prior.
- Here, the historical data is borrowed all together via the power prior.

Posteriors with the Power Priors and Decision Rule (continued)

- **A new variation of power prior**:

$$\pi(\boldsymbol{\theta}|D_0, \boldsymbol{a}_0) \propto \left((\sigma_c^2)^{-\frac{1}{2}} \exp\{-\frac{n_0(\bar{y}_0 - \mu_c)^2}{2\sigma_c^2}\}\right)^{a_{01}}$$
$$\times \left\{(S_0^2)^{\frac{n_0-3}{2}}(\sigma_c^2)^{-\frac{n_0-1}{2}} \exp\left[-\frac{(n_0-1)S_0^2}{2\sigma_c^2}\right]\right\}^{a_{02}} \pi_0(\boldsymbol{\theta}),$$

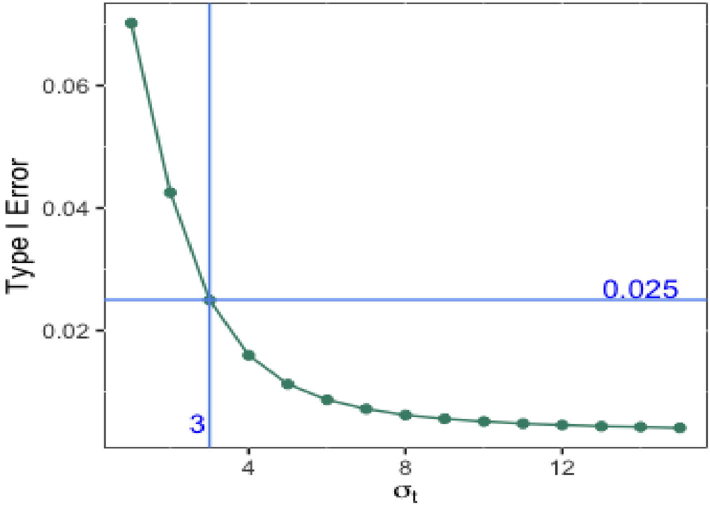
  where $(n_0, \bar{y}_0)$ and (n_0, S_0^2) are borrowed by parts with distinct discounting parameters a_{01} and a_{02}.

- $\pi_0(\boldsymbol{\theta}) \propto (\frac{1}{\sigma_c^2\sigma_t^2})^m$, where $m = 0$ corresponds to a uniform prior, $m = 1$ corresponds to a reference prior, and $m = \frac{3}{2}$ corresponds to Jeffreys's prior.

- Bayesian Decision Rule:

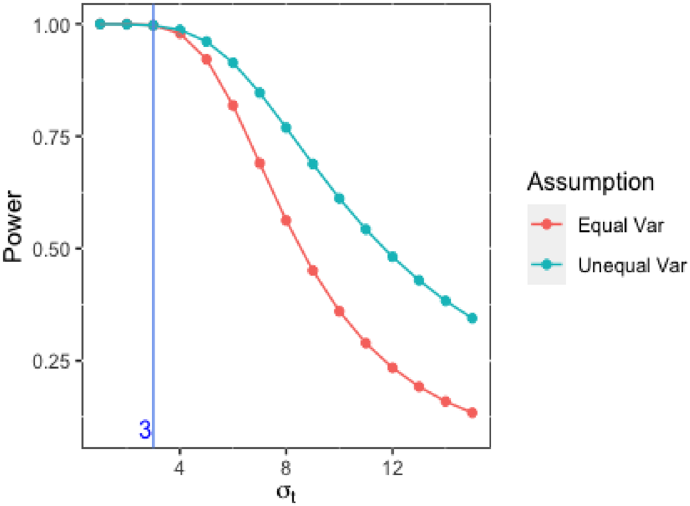
  Reject the null hypothesis of $\delta \leq 0$ if $P(\delta > 0|D_0, D) > \gamma$, where the credible level γ is chosen so that when $a_0 = 0$, the overall Type I error rate is intended to be controlled at 0.025.

Consequence of Assuming $\sigma_t^2 = \sigma_c^2$ on Type I Error ($n_t = 50,\ n_c = 25,\ \delta = 0$)



- When $\sigma_t^2 < \sigma_c^2$, Type I error is inflated. When $\sigma_t^2 > \sigma_c^2$, Type I error is deflated, leading to loss of power.

Consequence of Model Assumption on Power ($\sigma_t^2 = \sigma_c^2$?) ($n_t = 50,\ n_c = 25,\ \delta = 3.5$)

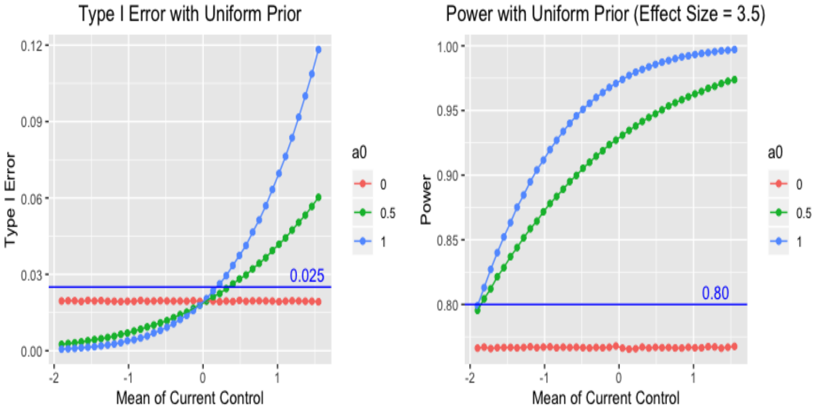


- When $\sigma_t^2 > \sigma_c^2$, the equal variance model leads to loss of power.

Effect of $\pi_0(\boldsymbol{\theta})$ on Bayesian Type I Error without Borrowing ($n_t = 50$, $n_c = 25$, and $\mu_c = 0$)

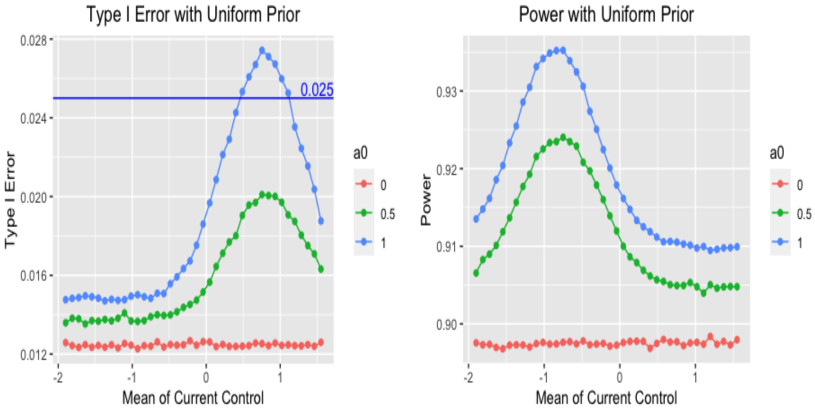
Bayesian Type I Error						
SD of Placebo	Assuming Different Sigma			Assuming Same Sigma		
	Mean of Current Control = 0					
	SD of Test			SD of Test		
	4	4.5	5	4	4.5	5
	Uniform Prior					
4	0.0192	0.0197	0.0192	0.0237	0.0192	0.0165
4.5	0.0193	0.0192	0.0193	0.0282	0.0236	0.0197
5	0.0195	0.0197	0.0198	0.0321	0.0277	0.0232
	1/sigma^2 Prior					
4	0.0235	0.0236	0.0233	0.0250	0.0207	0.0176
4.5	0.0240	0.0234	0.0234	0.0298	0.0249	0.0216
5	0.0239	0.0236	0.0232	0.0343	0.0293	0.0252
	Jeffrey's Prior					
4	0.0252	0.0252	0.0251	0.0257	0.0217	0.0181
4.5	0.0256	0.0253	0.0250	0.0304	0.0259	0.0220
5	0.0258	0.0255	0.0254	0.0354	0.0300	0.0257

Bayesian Type I Error and Power with Borrowing ($n_t = 50$, $n_c = 25$)



- Note: the maximum μ_c to control type I error is about 0.35 for $a_{01} = a_{02} = 0.5$, and is about 0.18 for $a_{01} = a_{02} = 1$.
- Most power gain is achieved by borrowing 50% of historical data.

Bayesian Type I Error and Power with Conditional Borrowing ($n_t = 50$, $n_c = 25$, $\sigma_c = 3.38$, $\sigma_t = 5$, $\bar{y}_0 = -0.18$, $S_0 = 3.38$)



- Note: $\pi_0(\boldsymbol{\theta})$ is the uniform prior, $\delta = 3.5$ for the power calculation, and the borrowing region of $0.5 \times SE$ for both mean and SD.
- The type I is much smaller.
- Again, most power gain is achieved by borrowing 50% of historical data.

Bayesian Type I Error and Power with Borrowing by parts ($n_t = 50$, $n_c = 25$, $\bar{y}_0 = -0.18$, $S_0 = 3.38$)

Bayesian Type I Error and Power with Unknown Variance					
a01	a02	Mean of Current Control = 0			
		Type I Error	Power		
			Effect Size = 3.5	Effect Size = 4	Effect Size = 4.5
Jeffrey's Prior, Sigma_t = 5, Sigma_c = 5, S_0 = 3.38					
0	0	0.0254	80.15%	89.44%	94.98%
0.5	0	0.0173	92.21%	97.22%	99.20%
1	0	0.0159	96.49%	99.10%	99.83%
0	0.5	0.0326	83.82%	91.78%	96.35%
0	1	0.0379	85.51%	92.81%	96.87%
0.5	0.5	0.0225	93.96%	98.02%	99.48%
1	1	0.0227	97.68%	99.46%	99.91%
Jeffrey's Prior, Sigma_t = 5, Sigma_c = 3.38, S_0 = 3.38					
0	0	0.0249	94.12%	98.08%	99.49%
0.5	0	0.0217	98.10%	99.61%	99.94%
1	0	0.0225	99.13%	99.86%	99.99%
0	0.5	0.0247	94.28%	98.15%	99.51%
0	1	0.0246	94.36%	98.19%	99.53%
0.5	0.5	0.0214	98.18%	99.64%	99.95%
1	1	0.0221	99.16%	99.88%	99.99%

Discussion

- Even with a "non-informative" prior, Bayesian type I error can be deflated, leading to loss of power.
- A mis-specified model may lead to a substantial inflation of type I error even under non-informative priors.
- Historical data can be borrowed by parts.
- Conditional borrowing or partial borrowing can further protect type I error.

Acknowledgement

I would like to thank all of my collaborators for their contributions on these 3 topics:

- Power Prior: Joseph G. Ibrahim (UNC), Yeongjin Gwon (UNMC), and Fang K. Chen (SAS)
- Information and Dissonance: Wei Shi, Lynn Kuo, and Paul Lewis (UConn)
- Bayesian Design: Wenlin Yuan (University of Connecticut) and John Zhong (Regenxbio).

Thank you !