Digital Twins & Bayesian Dynamic Borrowing:

two recent approaches for incorporating historical control data

Carl-Fredrik Burman, Erik Hermansson, David Bock,
Stefan Franzén, David Svensson

Statistical Innovation, Biometrics, Biopharmaceutical R&D,
AstraZeneca, Gothenburg, Sweden

PSI Amsterdam, June 2024

MAIN PAPER | Full Access

Digital twins and Bayesian dynamic borrowing: Two recent approaches for incorporating historical control data

Carl-Fredrik Burman ✉, Erik Hermansson, David Bock, Stefan Franzén, David Svensson

First published: 04 March 2024 | https://doi.org/10.1002/pst.2376

The presentation will omit details; please check the paper!

Gaining power from historical data

Background: Randomised clinical trial (RCT) comparing Active vs Control. Standard linear normal model for response y. Covariates $\mathbf{x} = \{x_1, x_2, \ldots\}$.

Two **very different** approaches to learn from Historical Controls (HC) data in order to gain efficiency for an upcoming RCT.

- Digital Twins (DT)
  - Machine Learning on HC
  - Find best predictor, $z(\mathbf{x})$.
  - Use z in the ANCOVA for the RCT
- Bayesian Dynamic Borrowing (BDB)
  - Borrowing: Use historical controls as controls in the RCT
  - Bayesian: Don't trust HCs completely
  - Dynamic: Let data tell you how much HCs can be trusted

Main focus on:

Bayesian Dynamic Borrowing

BDB: We want to estimate RCT treatment difference, δ

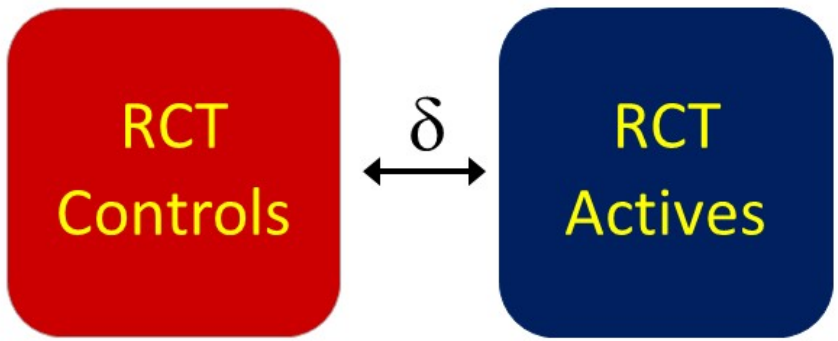


Figure: δ is mean treatment effect vs control in RCT

We think HC are similar to Trial Controls (TC)

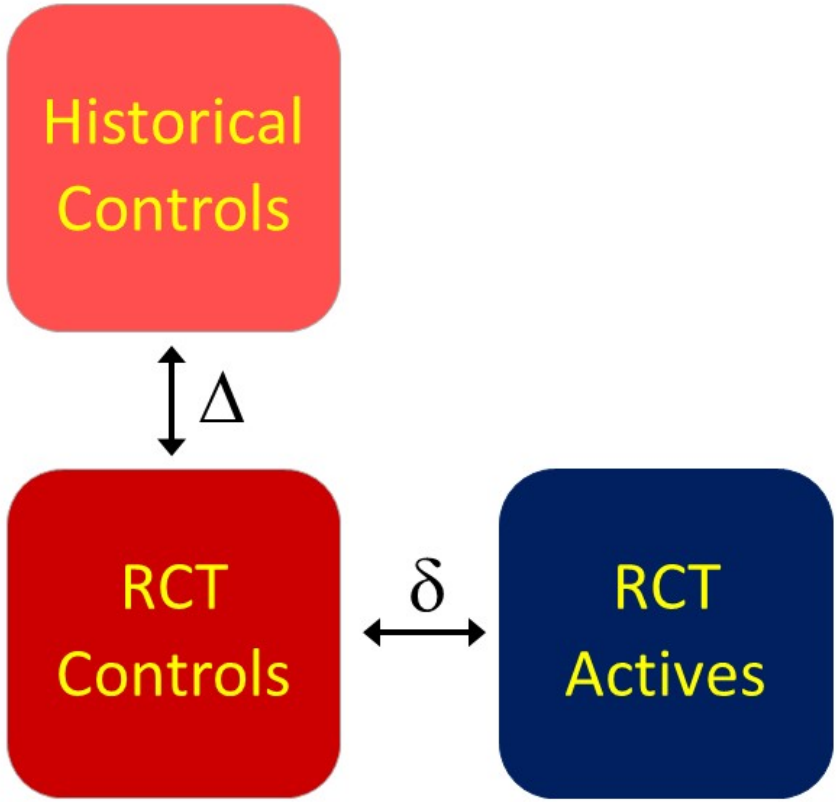


Figure: Δ is mean difference TC vs HC

Borrowing

Depending on the specific situation, historical controls can be expected to be more or less similar to trial controls. If we knew that HC and TC were essentially drawn from the same sample, we could just pool all controls and compare to patients on Active in the RCT. However, mean responses for controls could be different in historical and trial data.

Risk of bias if we compare RCT Active with HCs

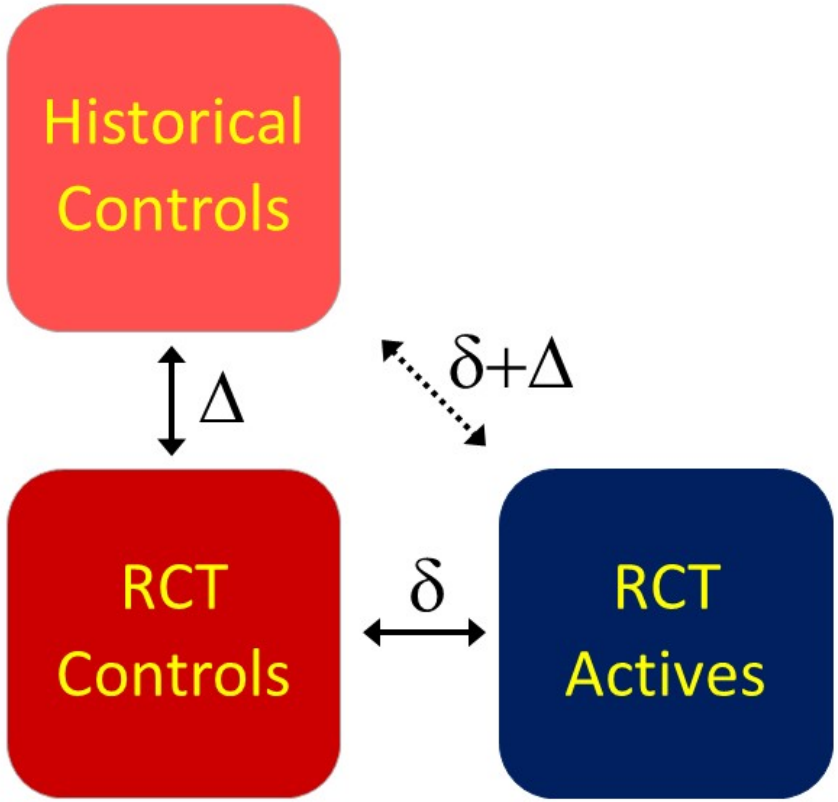


Figure: Bias Δ

For now, assume same sample size, residual variance

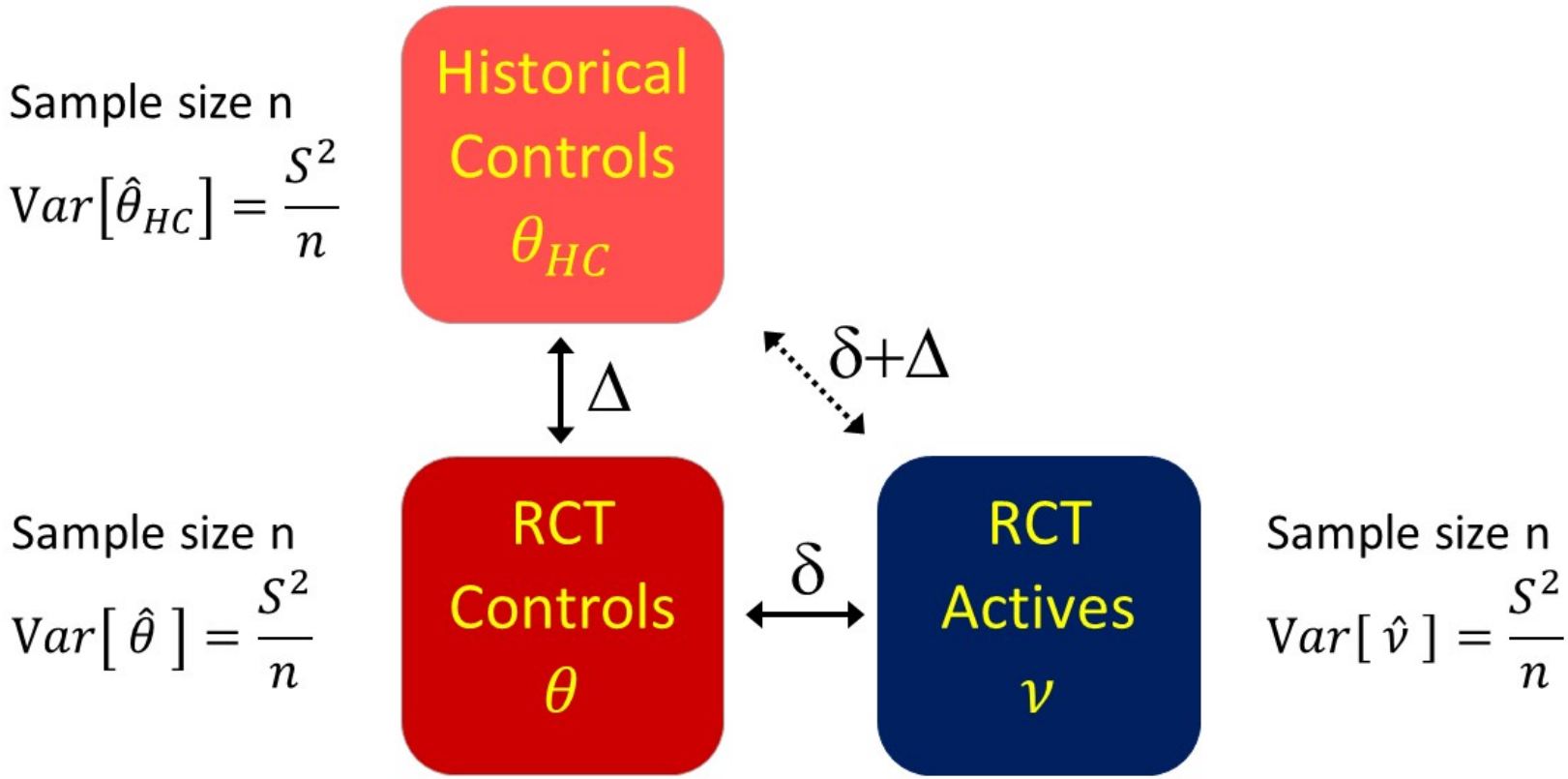


Figure: Sample size is n in each of the three groups. Same residual variance S^2. Sample variance per group: $\sigma^2 = S^2/n$.
$Var[\hat{\Delta}] = Var[\hat{\theta} - \hat{\theta}_{HC}] = 2\sigma^2$.

Bayesian Dynamic Borrowing: A simple variant

Optimistic scenario (No difference):
There is no difference in means for TC and HC, $\Delta = 0$.

Pessimistic scenario (vague prior):
We take "large" prior variance, $\Delta \sim N(0, \tau_\Delta^2)$ where $\tau_\Delta^2 \gg \sigma^2$.

Compromise (dynamic prior):
Prior weight $w = 0.5$ on both optimistic and pessimistic scenario.

Bayesian updating, given estimated difference $\hat{\Delta}$

Optimistic scenario (No difference):
There is no difference in means for TC and HC, $\Delta = 0$.
Prior is certain that $\Delta = 0$. Data cannot change this. Posterior also has $\Delta = 0$ with probability 1.
HC and TC are pooled.

Pessimistic scenario (vague prior):
We take "large" prior variance, $\Delta \sim N(0, \tau_\Delta^2)$ where $\tau_\Delta^2 \gg 2\sigma^2$.
Information in data overwhelms the info in the prior. The posterior for Δ is therefore approximately $N(\hat{\Delta}, 2\sigma^2)$.
HC are essentially disregarded.

Compromise (dynamic prior):
Prior weight $w = 0.5$ on both optimistic and pessimistic scenario.
We'll see later how the weight is updated.

Mixture prior for θ based on HC estimate $\hat{\theta}_{HC}$

Optimistic scenario (informative prior component):
There is no difference in means for TC and HC, $\Delta = 0$.
Leads to prior for θ *given* $\hat{\theta}_{HC}$: $\pi_0(\theta) = N(\hat{\theta}_{HC}, \sigma^2)$

Pessimistic scenario (vague prior component):
We take the prior information equal to the information for one single observation.
Prior for θ: $\pi_1(\theta) = N(\hat{\theta}_{HC}, S^2)$

Compromise (dynamic prior):
Prior weight $w_i = 0.5$ on both optimistic and pessimistic scenario.
Prior: $w_0 \cdot \varphi\left(\frac{\theta - \hat{\theta}_{HC}}{S}\right) + w_1 \cdot \varphi\left(\frac{\theta - \hat{\theta}_{HC}}{\sigma}\right)$

Updating normal priors

Normal priors are updated with normal data in a . . . *normal* way.

If only one normal prior:

- Assume prior for θ: $N(0, \tau^2)$. Data: estimate $\hat{\theta}$ is $N(\theta, \nu^2)$.
- Define prior *information* $\mathcal{I} = 1/\tau^2$ and information in data $\mathcal{J} = 1/\nu^2$.
- Prior mean and data estimate are simply weighted according to info. Posterior information is sum of prior info and data info. Thus, posterior is $N\left(\frac{\mathcal{I}\cdot 0 + \mathcal{J}\cdot\hat{\Delta}}{\mathcal{I}+\mathcal{J}}, \frac{1}{\mathcal{I}+\mathcal{J}}\right)$

Updating the mixture prior for θ given TC data, $\hat{\theta}$

Relax the assumption of equal sizes; let HC sample size be k times TC size, $n^\star = k \cdot n$. Still, $\sigma = S^2/n$ is TC variance.

Informative prior component (pool HC and TC)

Prior: $\pi_0(\theta) = N(\hat{\theta}_{HC}, \sigma^2)$

Posterior: $N\left((\frac{k}{k+1}\hat{\theta}_{HC} + \frac{1}{k+1}\hat{\theta}), \frac{\sigma^2}{k+1} \right)$

Vague prior component

Prior: $\pi_1(\theta) = N(\hat{\theta}_{HC}, S^2)$

Posterior: $N\left((\frac{1}{n+1}\hat{\theta}_{HC} + \frac{n}{n+1}\hat{\theta}), \frac{n}{n+1}\sigma^2 \right)$

Posterior weight The posterior weight ought to reflect whether HC are consistent with TC.

Let $s_i^2 = Var[\pi_i(\theta)] + \sigma^2$ be the predictive prior variance.

Posterior weights for informative/vague components are proportional to $\varphi\left(\frac{\hat{\theta} - \hat{\theta}_{HC}}{s_i}\right)$

Example of posterior weight

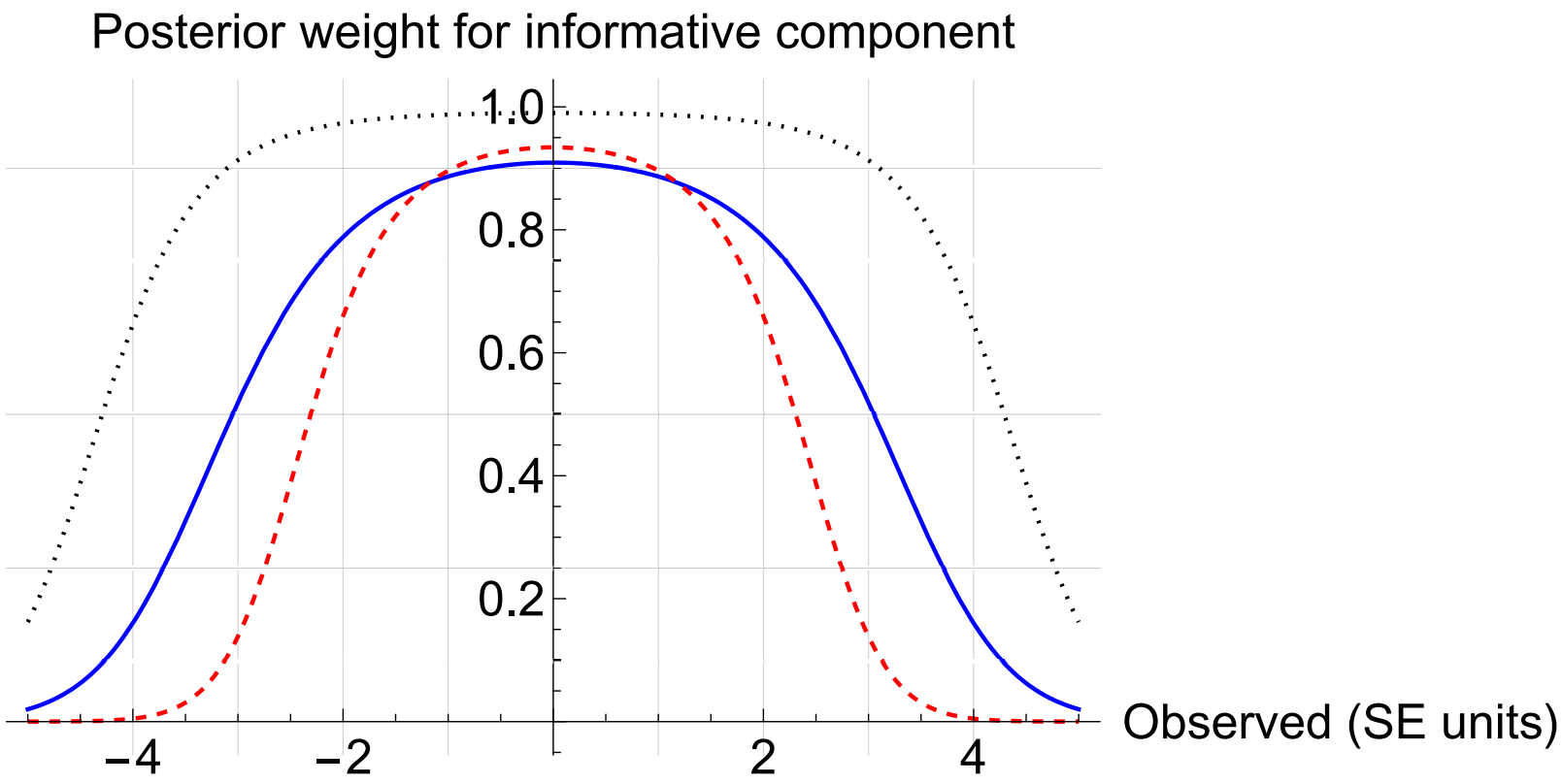


Figure: Weight for the informative component, as function of $\hat{\Delta}/SE[\hat{\Delta}]$. Solid blue curve has $n^\star = n = 200$. The dotted black curve results when the vague prior has ten times larger standard deviation. Finally, the red dashed curve use the same model as the blue but with $n^* = \infty$.

Posterior weight is pretty random!

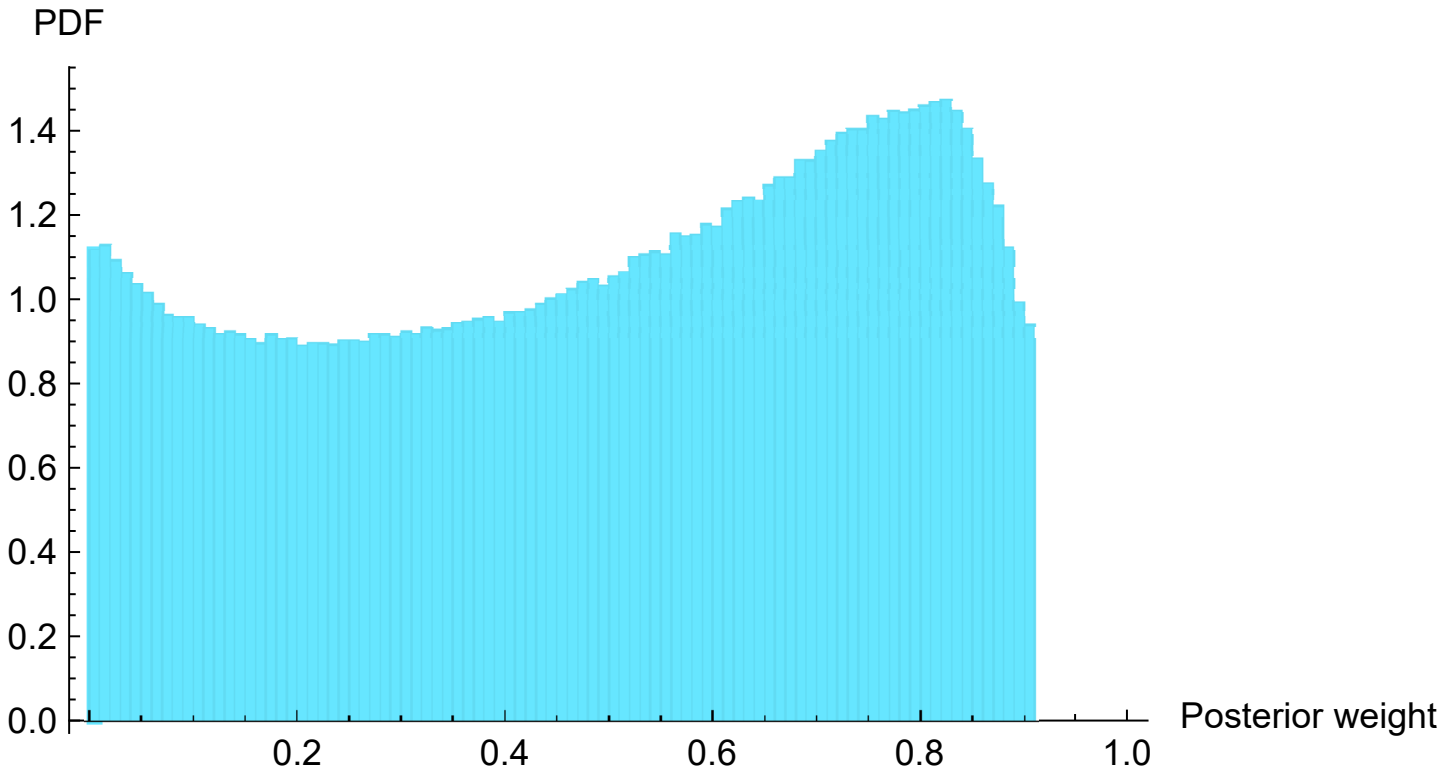


Figure: Probability density for posterior weight when RCT control mean is 3 standard error units away from HC average.

We cannot solve the problem statistically

We don't know if HCs are perfect or useless.
We only get *one* estimate of HC vs RCT difference.
This randomly tells us what to think. NOT GOOD!

If we have very good power, we don't need BDB
If we have low/moderate power, then $\hat{\Delta}$ will automatically be uncertain in any situation we'd be interested in.

Conclusion: We need extra-statistical information. Real clinical insight and common sense, not a mathematical black box!

Very large Type 1 Error inflation

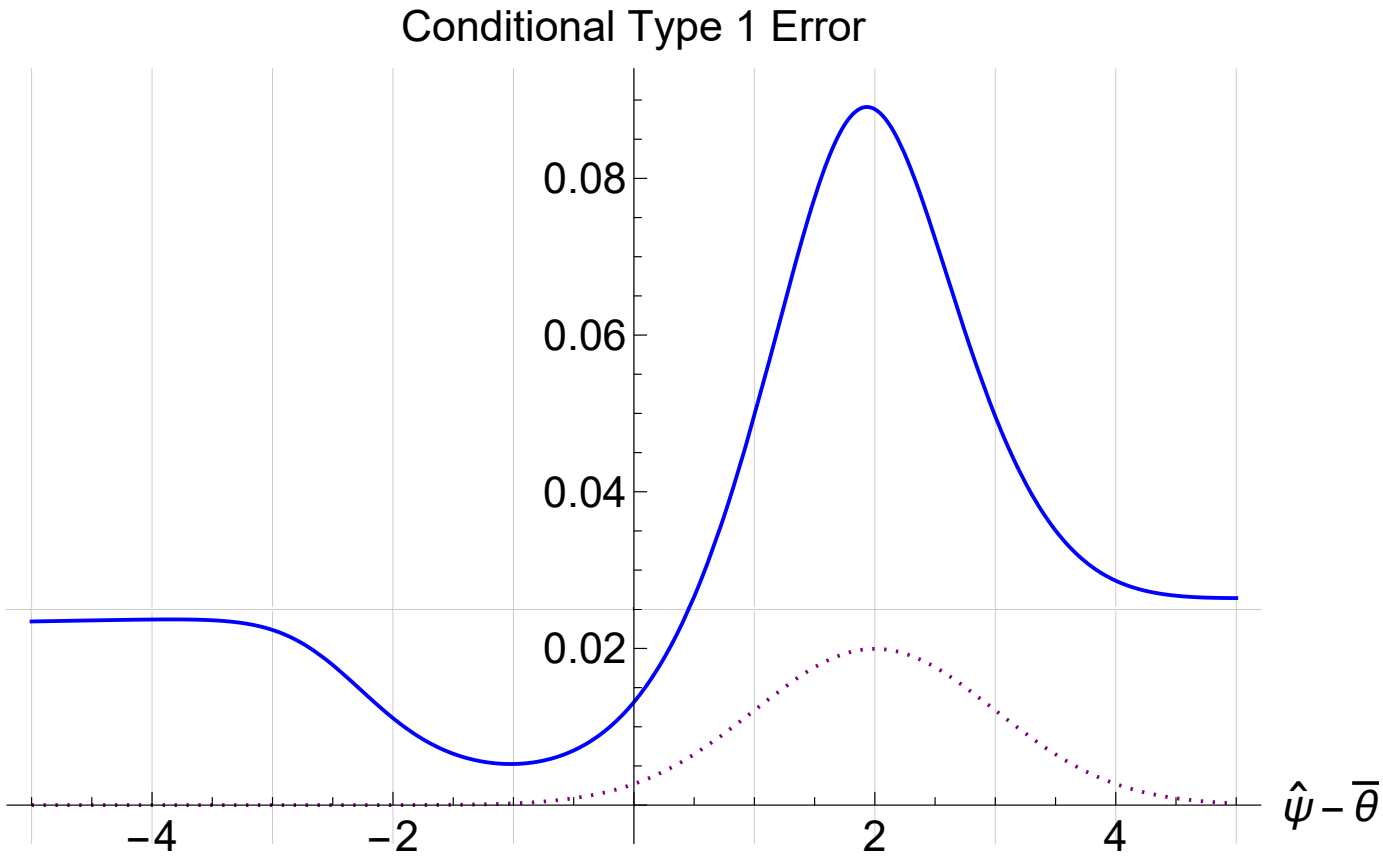


Figure: Conditional Type 1 Error given both the average (Active and Control) observed RCT response, $\hat{\psi}$, and the observed Historical Control average, $\bar{\theta}$. Variables are scaled so that the standard error of $\hat{\psi}$ is one. The dotted curve indicate the shape (scaled to fit the figure) of the normal density for $\hat{\psi}$ if $\psi = 2$.

What (not) to condition on

Standard (frequentist) inference theory:

- You should condition on nuisance parameters! ("Nuisance" parameters do not tell us anything about the relative treatment effect, in the model we have.)

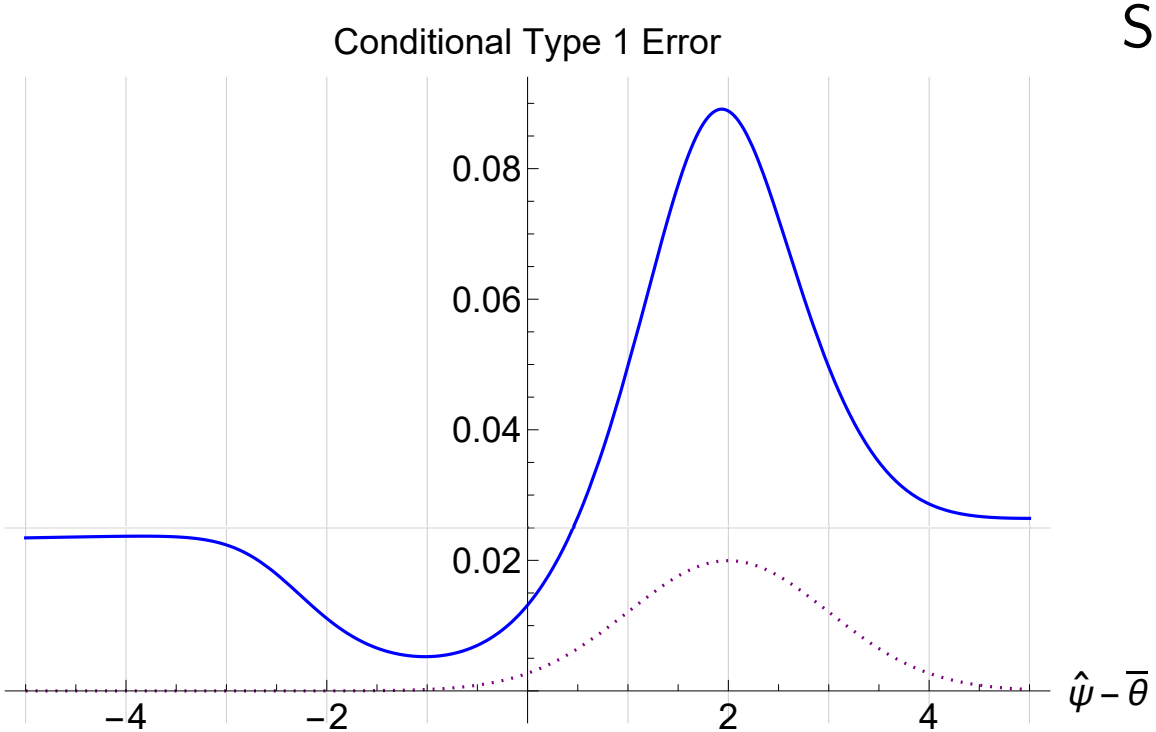
In the analysis of a standard non-complex RCT, this means e.g.

- Take the sample sizes for each arm as fixed (although total sample size is random due to when recruitment stops, split depends on randomization).
- In survival analysis, take the total number of events as fixed together with inclusion times.
- In normal model, take response of *all* RCT patients as fixed.

BDB uses a "larger" model. Conditioning is less obvious!

Conditional T1E revisited

Different (conditional) Type 1 Errors may be relevant for different scenarios.



Scenarios:

1. Look at blinded RCT data, *then* choose HCs
2. Estimate HC average, then run RCT.
3. Blind both HC and RCT data.

Real Bayesian instead of black box

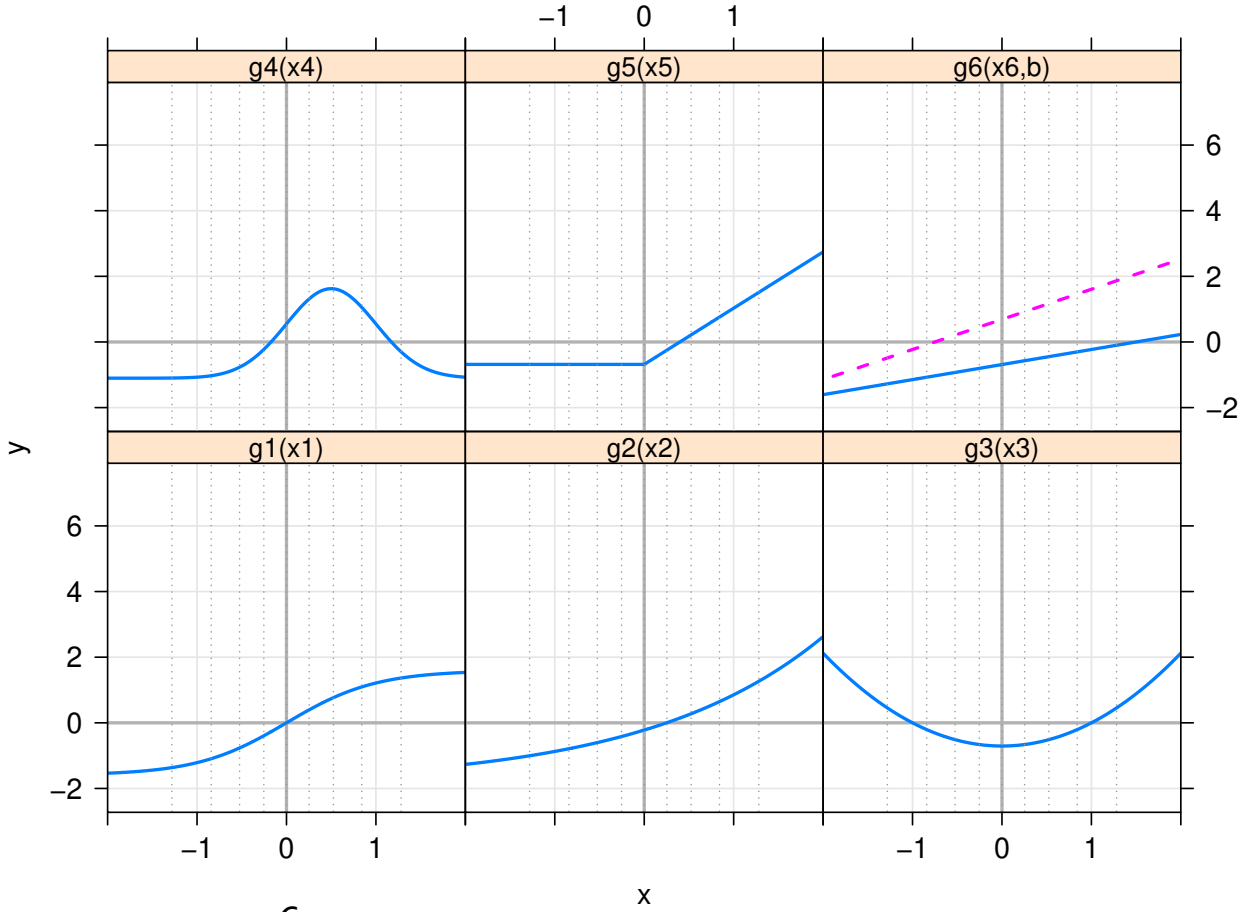
What do we (experts) believe? (RCT vs RCT in same centers with same sponsor; RCT vs RCT with other sponsor; RCT vs Register data)
Compare covariate distributions.
Often better responses over time? Often better responses in RCT?
Prior with non-zero mean drift. Prior for between-trial variability τ.

Digital Twins

DT simulation set-up

$$y(\mathbf{x}, \mu) = \mu + g_1(x_1) + g_2(x_2) + g_3(x_3) + g_4(x_4) + g_5(x_5) + g_6(x_6, b) + \varepsilon$$

where $x_j \sim N(0, 1)$ independent



$$\mathbf{Var}(y) = \sum_{j=1}^{6} \mathbf{Var}(g_j(\mathbf{x})) + \mathbf{Var}(\varepsilon) = 6 \cdot 1 + 4 = 10.$$

Variance explained by Random Forest or Linear Model

			Variance explained			
		Variance	LM	R50	RF200	RF1000
Total		**10**	**3.21**	**1.91**	**3.49**	**4.56**
Logistic	$g_1(x_1)$	1	0.93	0.56	0.72	0.75
Exponential	$g_2(x_2)$	1	0.88	0.48	0.65	0.75
Quadratic	$g_3(x_3)$	1	0.00	0.19	0.37	0.55
Bell	$g_4(x_4)$	1	0.20	0.28	0.48	0.68
Trend change	$g_5(x_5)$	1	0.73	0.48	0.68	0.78
Interaction	$g_6(x_6, b)$	1	0.47	0.20	0.30	0.36
Residual	ε	4				

Table: Variance / variance explained, in total and separately for the six model components. For Random Forest, $n^* = 50, 200, 1000$. The values presented for each variable x_j are a difference between the overall MSE and the MSE resulting if deleting the variable x_j from the data set, and then re-fitting the model. May not add to total!

DT simulation results

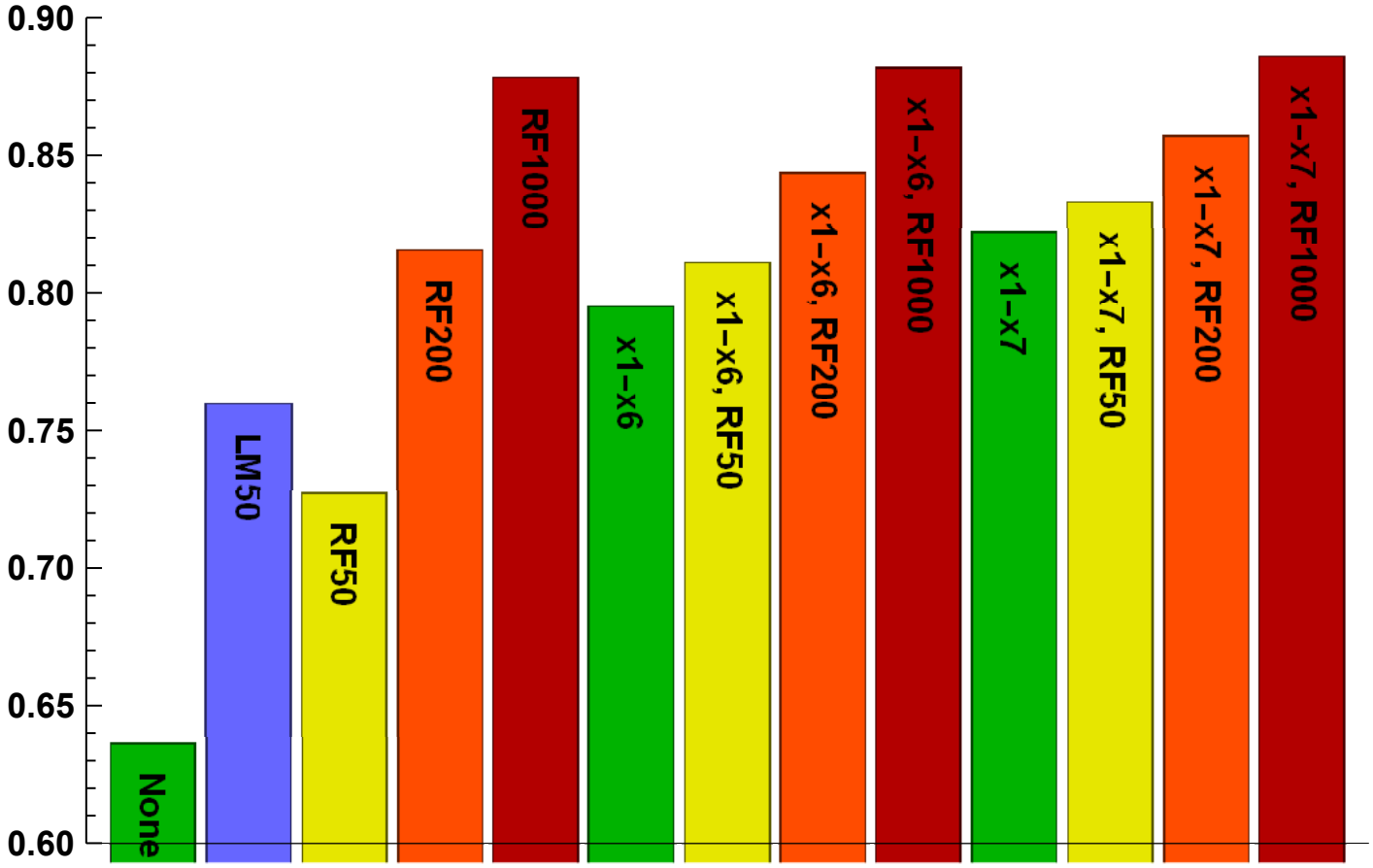


Figure: Power for ANCOVA models with 0, 6 or 7 baseline covariates (green); similar models complemented by an RF-trained predictor, based on $n^* = 50$ (yellow), $n^* = 200$ (orange) or $n^* = 1000$ (dark red); or complemented by a linear predictor ($n^* = 50$, blue) based on historical data.

DT discussion

- Learn predictor from HC
- Include in ANCOVA
- Potential gain, almost no statistical cost
- How large non-linearities in real applications? Need more practical examples.
- Learn from HC! But perhaps enough to add simple covariates?

- Digital Twins and Bayesian Dynamic Borrowing use historical control data in very different ways