The Use of Historical Information in Clinical Trials

Kert Viele

CTMC 2017

Acknowledgements

- DIA BSWG coauthors on manuscript (appeared in Pharmaceutical Statistics)
  - Scott Berry, Beat Neuenschwander, Billy Amzal, Fang Chen, Nathan Enas, Brian Hobbs, Joe Ibrahim, Nelson Kinnersley, Stacy Lindborg, Sandrine Micallef, Satrajit Roychodhury, Laura Thompson
- Other helpful discussions
  - Jeff Wetherington, Linda Mundy

Introduction

- Many trials compare a novel treatment to a control arm.
- Control rarely exists in vacuum
  - many studies on control effectiveness
  - trials used for approval
  - trials post approval
  - etc.
- Prior to the study, we often believe we have a "good idea" of the control arm parameters.
- Can we use this information?

Historical Borrowing versus Historical Control

- "Historical Control" often refers to single arm (e.g. enroll NO controls)
  - often compare to a single number based on the historical data (e.g. "beat 20%")
  - if correct, most statistically efficient
  - if your number is wrong, no way to EVER know
- "Historical Borrowing" refers to augmenting a randomized trial
  - have control arm, perhaps with unequal randomization (2:1, 3:1, etc)
  - control arm inferences combine current and historical data
  - far less risk, some less benefit

Which studies to borrow from? (big deal, not the focus here)

- Is the historical information "on point"
- Very big deal, not covered here
  - often studies found by literature search
    - are they representative?
    - certainly shouldn't be "cherry picked"
  - can find patient records at sites prior to study commencement
  - platform trials

Simple Example

- Dichotomous endpoint
- Current trial has 200 subjects on each of control and treatment arms
  - $Y_C \sim Bin(200,p_C)$ $Y_T \sim Bin(200,p_T)$
- Historical data available with 65/100 responses. ($Y_H \sim Bin(100,p_H)$)
- Primary Analysis
  - $H_0 : p_C=p_T$ versus $H_1 : p_C < p_T$

Basic idea of historical borrowing

- Combine information from current and historical study
  - typically through informative prior, frequentist methods also possible
- Can be good, can be bad
  - if historical information is close to true current parameters, inferences much improved
  - if historical information is far from true current parameters, bias can occur
  - discrepancy between history/current we call drift
  - unfortunately, drift not known in advance

Causes of Drift

True parameters can change from study to study

SOC changes over time
(maybe better, maybe worse)
SOC can qualitative change (new drug added to SOC)
Different sites

Different population (BEWARE CHERRY PICKING HISTORICAL DATA!!!)

Main differences are pretty well understood

Sampling variability we would encounter in any study.

BEWARE CHERRY PICKING HISTORICAL DATA!!!

Downweighting/Power priors

- Weight historical data relative to current
- Each historical subject "counts" as W current subjects
  - W=0 ignores historical data
  - W=1 corresponds to pooling
  - W<1 "downweights" historical subjects
  - W>1 "overweights" historical subjects (rarely done)
  - W=infinity is a single arm trial…

Fixed Weights

- Place noninformative priors on p_C and p_T.
- Use likelihood for control arm of
  - $[p_C^{65} (1-p_C)^{35}]^W \quad [p_C^{YC} (1-p_C)^{(200-YC)}]$
  - W = weight of historical data
- Example W=0.2, the 65/100 acts like 13/20
- Consider W=0.0, 0.2, 0.4, 0.6, 0.8, 1.0
- Borrow "equivalent" of 100W subjects

Fixed Weights

- Posterior mean for p_C is
  - $g(Y_C) = (65W+Y_C) / (100W + 200)$

- Suppose you observed 146/200=73% responses on the control arm.
  - With W=0, posterior mean is 73%
  - With W=0.2, posterior mean is 72.27%
  - With W=1, posterior mean is 70.33% (pooled)
  - With W=100, posterior mean is 65.16%

Operating Characteristics

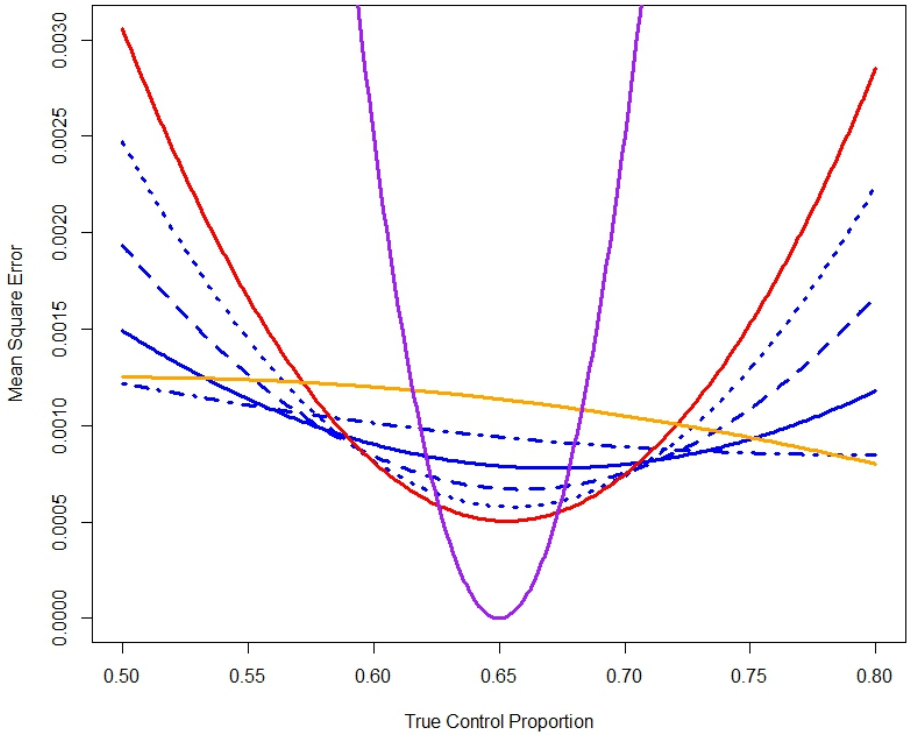
- Obtain the posterior distribution of p_C.
- Posterior mean $g(Y_C)$ on previous slide
  - compute MSE of point estimate
  - $MSE(p_C) = E[(g(Y_C) - p_C)^2 \mid p_C]$
- Also perform hypothesis testing
  - Reject H_0 if $Pr(p_T > p_C) > 0.975$
  - compute type I error $Pr(\text{reject} \mid p_T = p_C)$
  - compute power $Pr(\text{reject} \mid p_T = p_C + 0.12)$
  - power and type I error also a function of p_C

Fixed Weights (MSE as function of drift)

W=0 orange
W=1 red

single arm trial in purple

other weights in blue



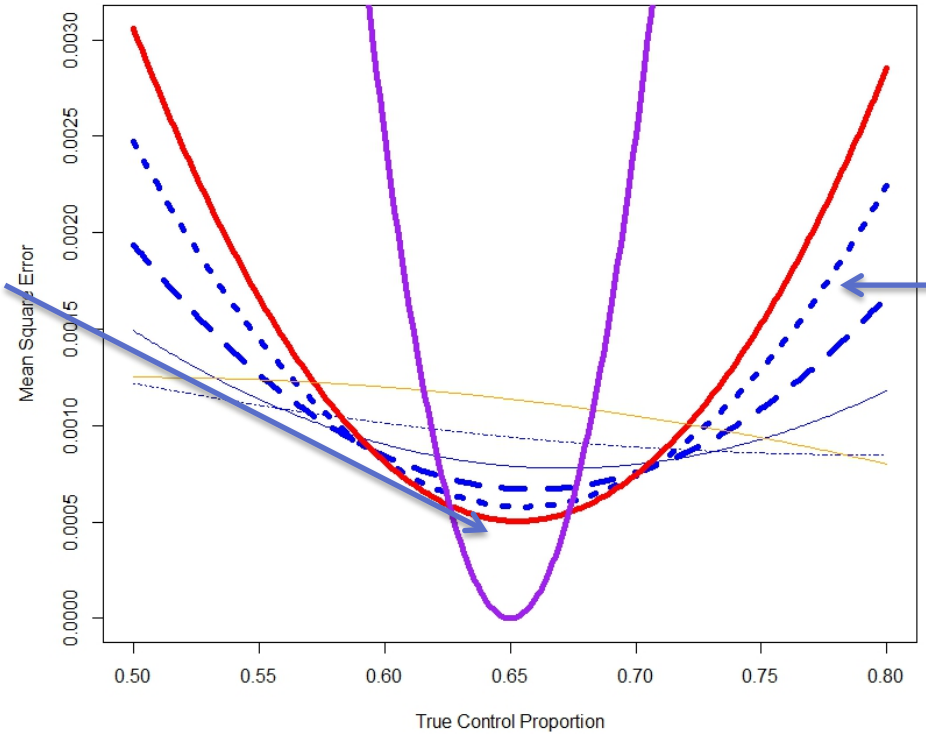
X-axis is p_C (current control Parameter)

Y-axis is MSE

If we knew drift, could select an ideal weight (of course we don't)

Fixed High Weights (MSE)

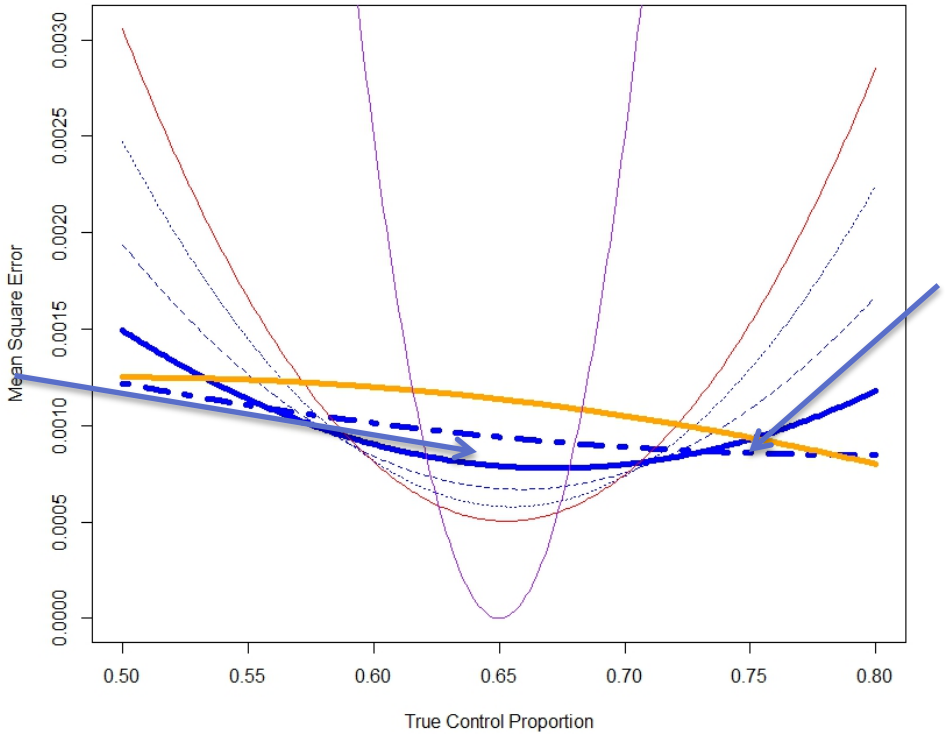
High Weight and No drift provides dramatic gains.



High drift and high weight produce biases and poor MSE

Fixed Low Weights (MSE)

Low weights and low drift produce more modest gains (compared to ignoring history)



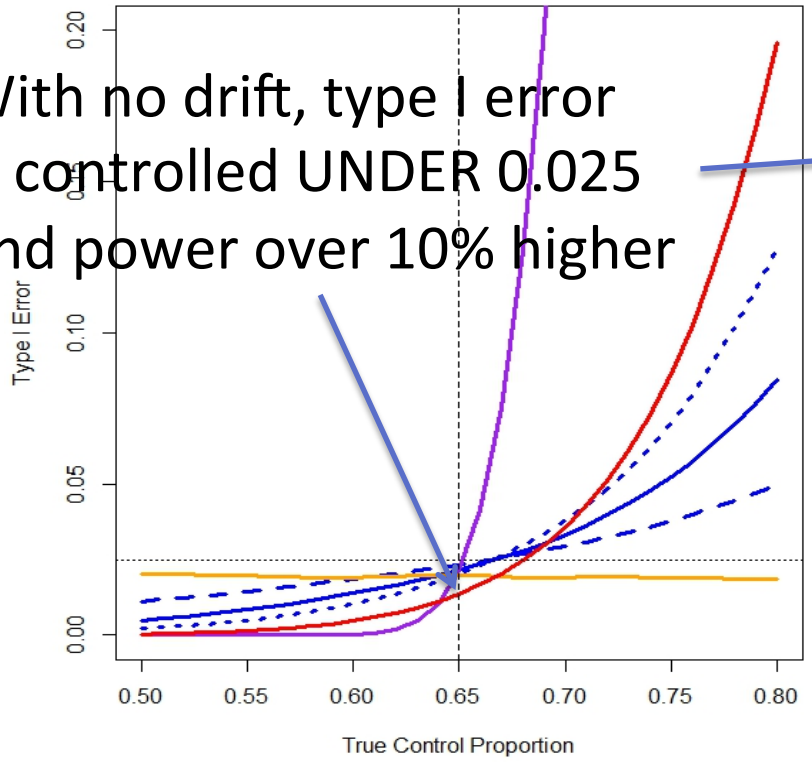
Low weight and High drift produce gains over broader area

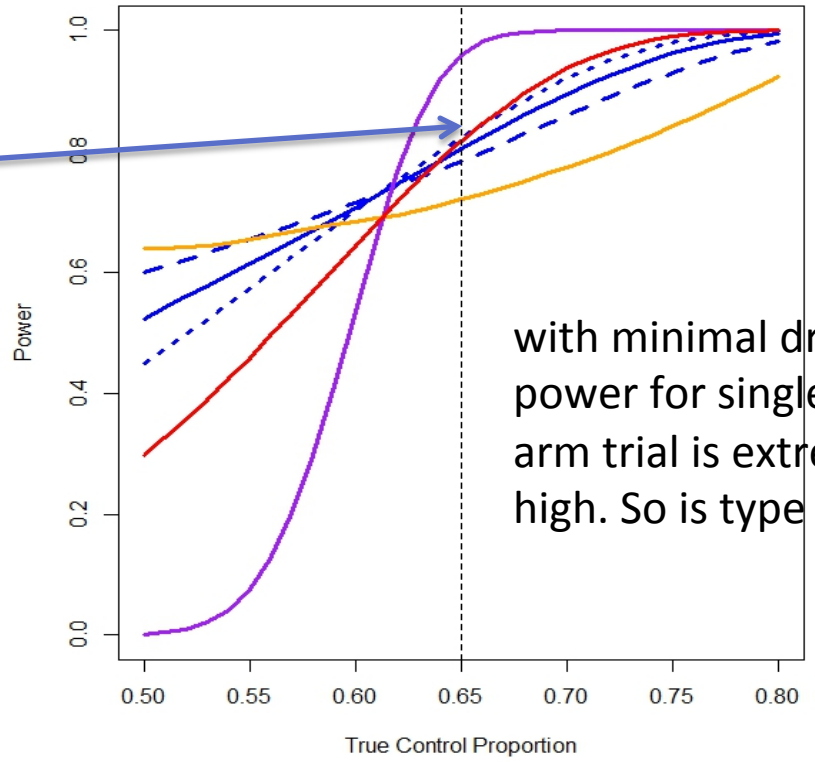
Fixed Weights (Testing)

Type I error

Power (for 0.12 gain)

With no drift, type I error is controlled UNDER 0.025 and power over 10% higher

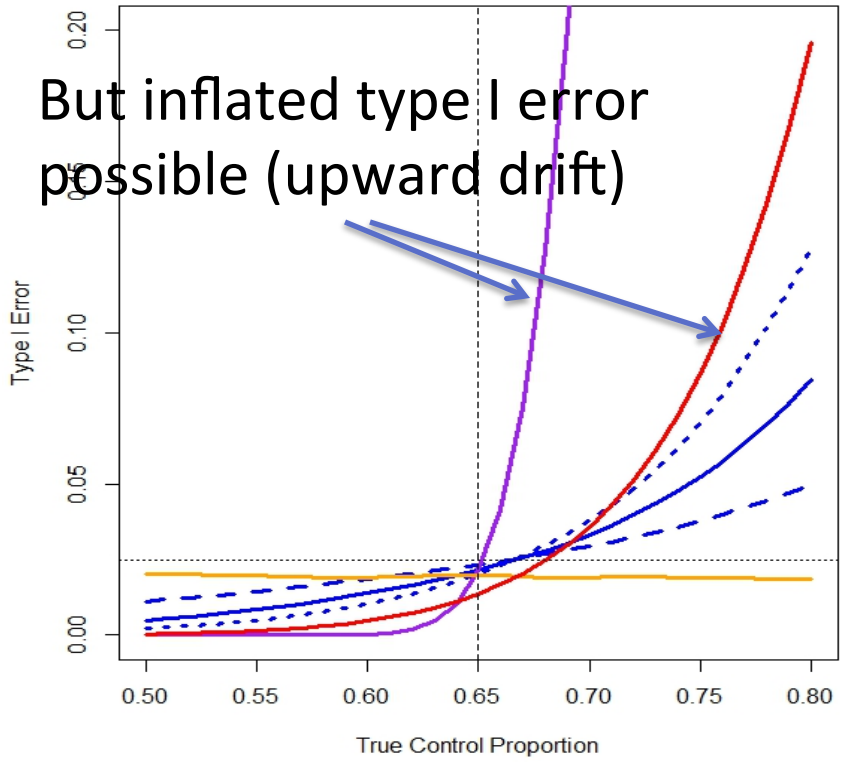




with minimal drift power for single arm trial is extremely high. So is type I error

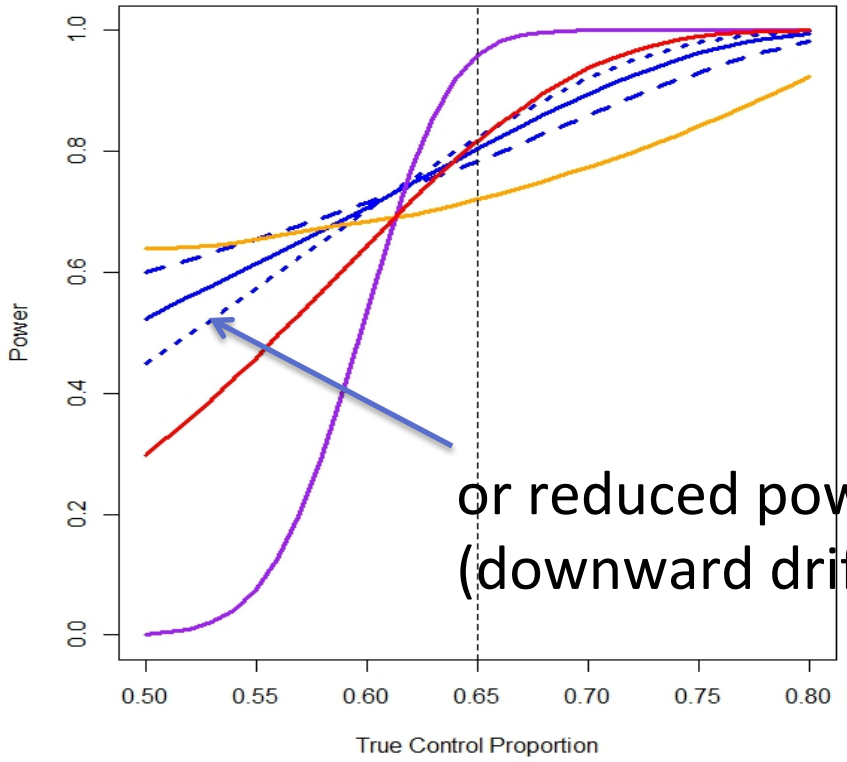
Fixed Weights (Testing)

Type I error



But inflated type I error possible (upward drift)

Power (for 0.12 gain)



or reduced power (downward drift)

Key point

- "Drift" – difference between observed historical and current data parameter
  - e.g. difference between 0.65 and p_C
- If we knew drift, we'd know how much to weight!
  - no drift, then use a large weight
  - lots of drift, then use a small weight
  - some drift, weight between 0% and 100%
- UGH! We never know drift in advance…

If only…

- "Those who cannot remember the past are condemned to repeat it."
  - George Santayana

- We are worried about the complete opposite…that history will not repeat…

Dynamic Borrowing

- Desired weight depends on unknown drift
  - small drift = large weight
  - large drift = small weight
- The data itself provides information on drift

- Dynamic borrowing = amount of weight depends on agreement between historical and current data.

Hierarchical Models (not the only way)

- In general, let p_C be current control rate
- $p_1,\ldots,p_H$ are true rates from historical studies
- $Y_0 \sim \text{Bin}(n_0, p_C)$ [current data]
- $Y_h \sim \text{Bin}(n_h, p_h)$ [historical data]
- $\text{logit}(p_C),\ldots,\text{logit}(p_H) \sim N(\mu,\tau)$
- $\mu \sim N(\mu_0,\tau_0)$, $\tau \sim \pi(\tau)$

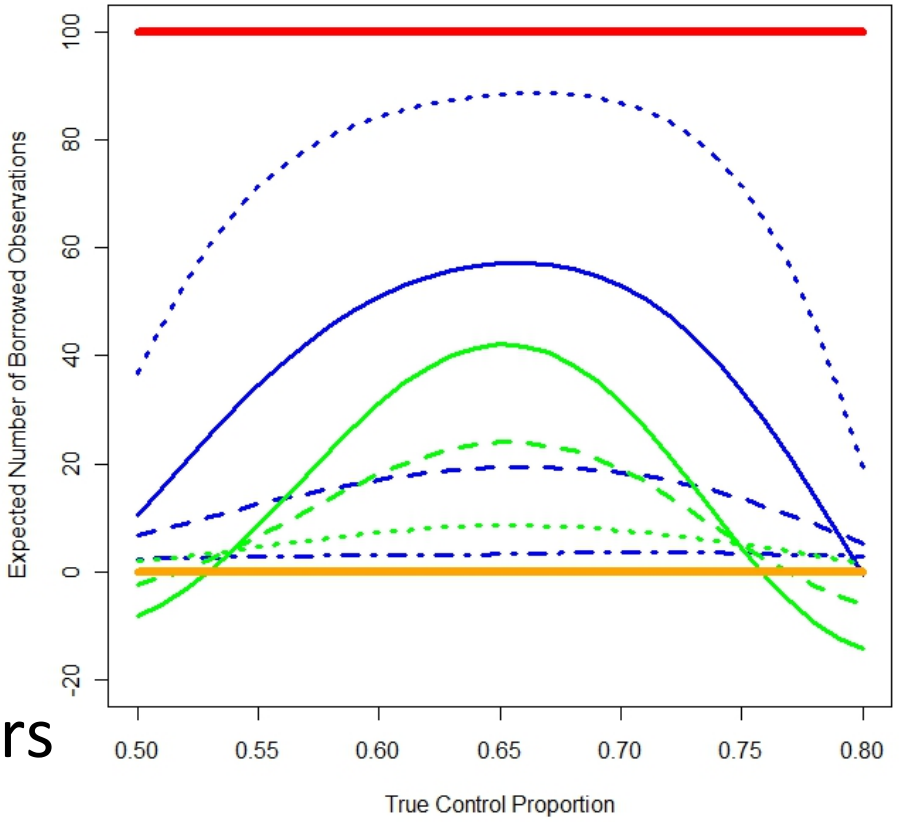
Hierarchical Models

- $\text{logit}(p_C),\ldots,\text{logit}(p_H) \sim N(\mu,\tau)$
- τ measures across study variation
- A fixed τ corresponds to a specific weight
- We use an IGamma prior, here we obtained good operating characteristics.
  - other prior structures available
- Creates dynamic borrowing
  - generally lower τ when current data agrees with history, and thus higher weight
  - generally larger τ when current data disagrees with history, and thus lower weight.

Hierarchical Models (expected borrowing behavior)

Y-axis shows expected number of borrowed subjects

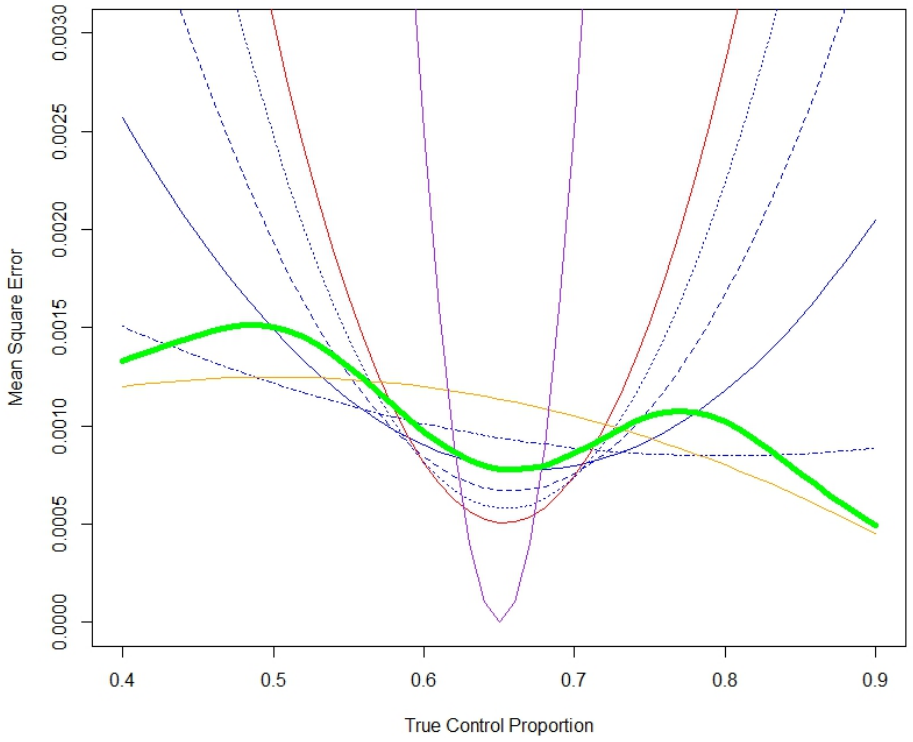
Different curves are different priors



Dynamic Borrowing - E[borrow] greatest for low drift

MSE for dynamic borrowing

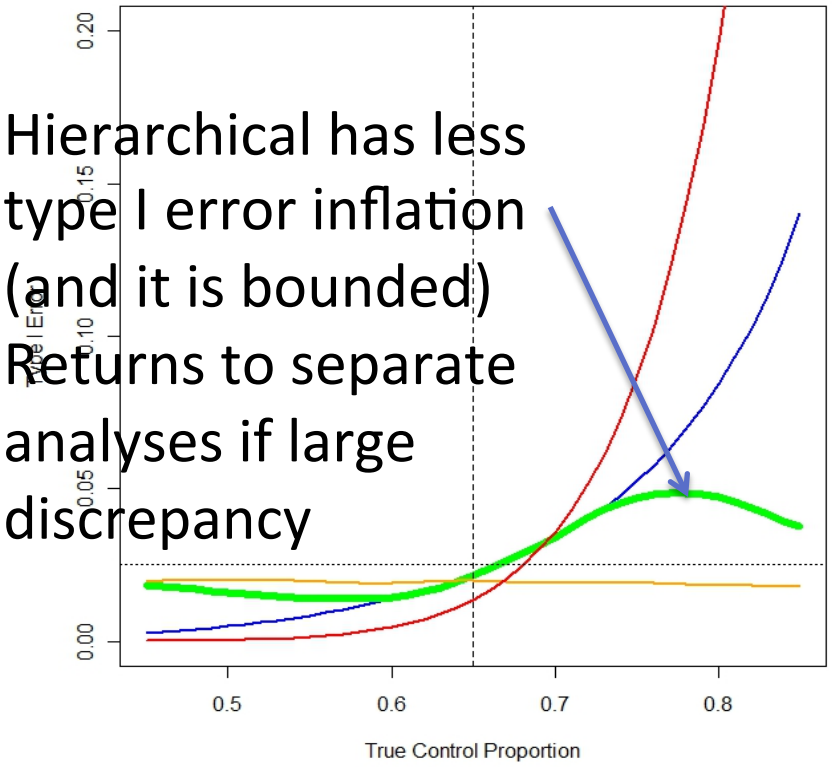
Green curve shows MSE for posterior mean using dynamic borrowing



Note inflation of MSE is bounded over ignoring history

Hierarchical Models

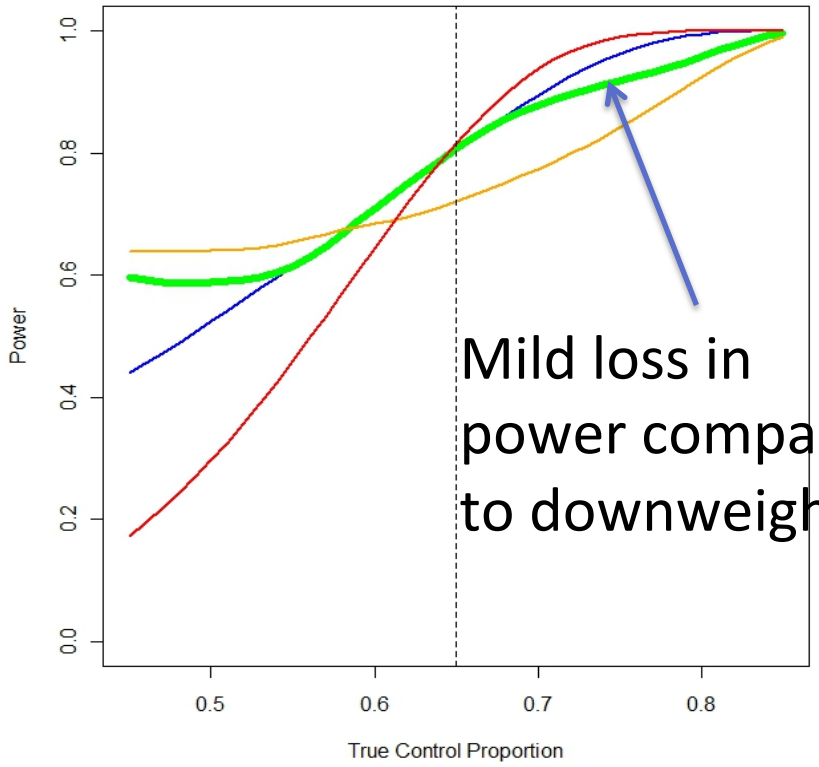
Type 1 Error



Hierarchical has less type I error inflation (and it is bounded)

Returns to separate analyses if large discrepancy

Power (0.12 gain)



Mild loss in power compared to downweighting

Example of Benefit of Design

- Original design noninferiority trial in antibiotics, required 750 subjects, 375 per arm.
- With historical borrowing (2 historical studies)
  - required 600 subjects (20% fewer)
  - randomized 200 (ctrl), 400 (trmt)
  - for “expected” drift, control of type I error and comparable power.

Comments on type I error

- Usual definition of type I error conditions on historical data
  - $\alpha(p_C) = Pr(\text{success} \mid p_C=p_T, Y_H)$
  - regardless of borrowing method type I error is inflated for SOME p_C (those with large drift)
- You could argue strictly this precludes historical borrowing.
  - However, this can lead to some unintuitive decisions.

Comments on type I error

- Note "drift" in this definition and in the OCs refers to p_C-0.65
  - difference between current control parameter and historical data
- Suppose you KNEW $p_H=p_C$
  - e.g. current control parameter equal to historical control parameter
  - this does NOT guarantee 0 drift because sampling variability still present.

Surely you must pool…

- In your favorite computer program…
- pc=[censored]
- pt=[censored]
- rbinom(1,10000,pc) [Y~Bin(10000,pc)]
  - result is 6399
- Now, **with no changes to pc and pt**, you are asked to design an experiment to test $H_0 : p_C = p_T$ with 200 observations per arm.

Surely you must pool…

- Can you use the 6399/10000 historical data?
- All the prior graphs still apply!
  - If $p_c=p_t=0.75$, type I error will be inflated.
- But you KNOW you are using the same p_c.
- You must pool! Can't ignore 10,000 observations
  - $p_c=0.75$ pretty unlikely given 6399/10000.

What is (currently) achievable?

- You cannot
  - dominate “ignore history” (ignore history wins if drift large enough)
  - globally decrease type I error and increase power

- You can
  - control maximal type I error inflation
  - control range where borrowing=improvement
  - get improvement for any fixed dist’n on drift

Summary

- Historical borrowing
  - may improve point estimates
  - may reduce type I error
  - may increase power
  - can result in substantial sample size savings
- There will be situations where historical borrowing is NOT beneficial
  - large expected drift, or high variation in drift