The Use of Historical Information in Clinical Trials

Kert Viele

Midwest Biopharmaceutical Statistics Workshop

2014

Acknowledgements

- 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

- Almost ALWAYS have historical information available
  - on the control/comparator arm (focus here)
  - on the current treatment
  - on similar drugs in the class
- How do we use this?
  - here focus on augmenting control information

Introduction

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

Introduction

- Here we assume the studies have been picked "well"
  - similar inclusion/exclusion criteria
  - similar sites
  - reasonably recent in time
- However, no two studies are identical
  - always going to be "drift"
    - patient populations don't exactly match
    - rates/means change over time
    - sampling variability

Introduction

- Will the historical data lead us to better conclusions?
  - basic answer is “it depends on drift”
  - small (can be nonzero!) drift results in improved operating characteristics
    - lower MSE, lower type I error rates, more power
  - large drift results in poorer operating characteristics
    - bias, inflated type I error OR power loss
- BUT……drift is unknown!

If only…

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

Outline of remainder of talk

- Discuss three levels of historical borrowing
  - What does it do with the data in a single trial?
  - What are the operating characteristics, fixing the parameters and integrating over the data
    - MSE, type I error, power
  - What are the long run properties of using historical data?
    - Under what conditions will we generally get better answers by using historical information.
    - This is function of the “drifts” we will see

Outline of remainder of talk

- Focus is on scientific conclusions
- Historical borrowing is quite technically developed
  - time varying covariates, etc.
- Acceptable for early phase trials
- Less accepted for confirmatory trials

Bayesian or Frequentist?

- This talk is largely Bayesian, but certainly can pursue a frequentist paradigm.
- Of course, Bayesian decision rules have frequentist properties
  - point estimates have bias, variance, MSE
  - hypothesis decisions have type I error, power

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 $H_1 : p_C < p_T$

Main idea

- For both estimation and testing, combine current and historical data on control

- How much weight to place on history?
  - fixed weight (0%=none, 100%=pool, in between is “downweighting”)
  - dynamic weight = weight based on agreement between current and historical data.

Key point

- "Drift" – difference between observed historical and current data
- If we knew drift, we'd know how much to weight!
  - no drift, then pool (100% weight to history)
  - lots of drift, then use a small weight
  - some drift, weight between 0% and 100%

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 in [0,1], weight of historical data (0=ignore, 1=pooled)
- 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
  - $(65W+Y_C) / (100W + 200)$

$$= \left[\frac{100W}{100W+200}\right]\frac{65}{100} + \left[\frac{200}{100W+200}\right]\frac{Y_C}{200}$$

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

Operating characteristics

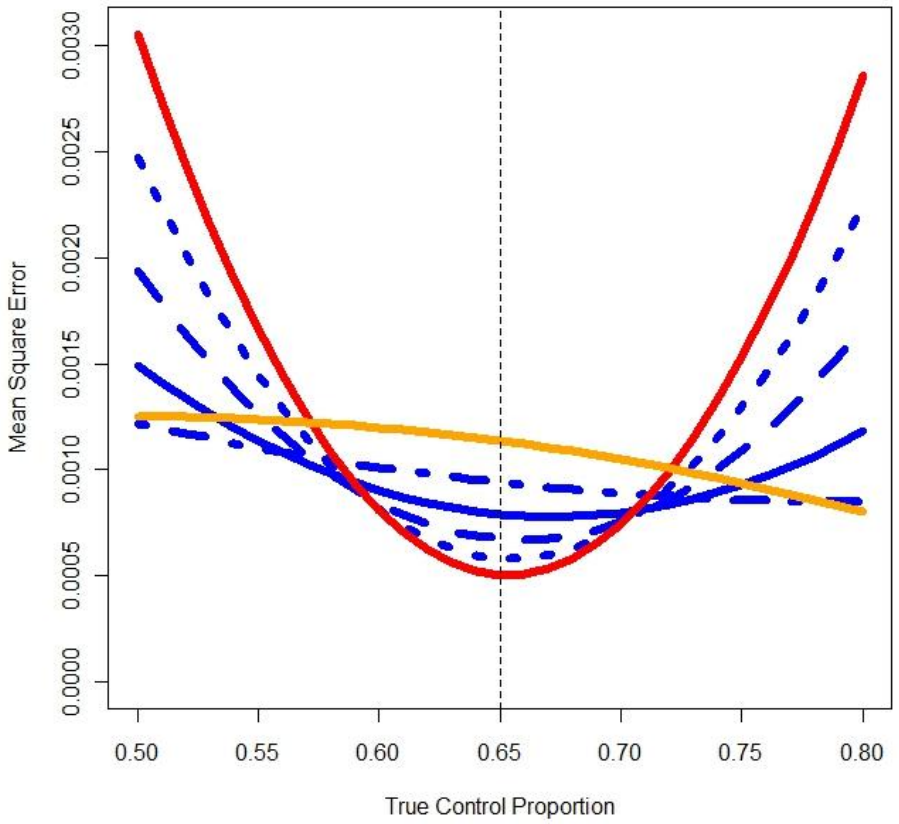
- For a fixed value of p_C, we know the distribution of the data.
  - $Y_C \sim \text{Binom}(n=200, p=p_C)$
- For each value of Y_C, we can find the posterior mean and the square error.
- We can then average to find mean square error as a function of p_C.

Fixed Weights (MSE as function of drift)

Ignoring historical data in orange (W=0)

Pooling in red (W=1)

other weights in blue



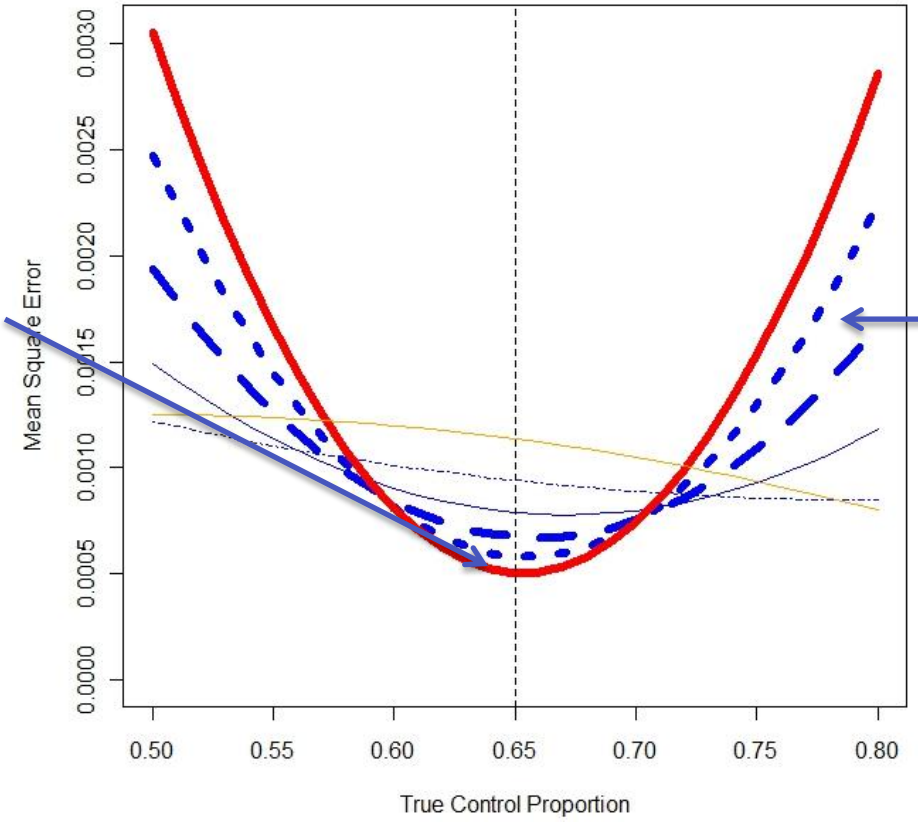
X-axis is p_C (current control Parameter)

Y-axis is MSE

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

Fixed High Weights (MSE)

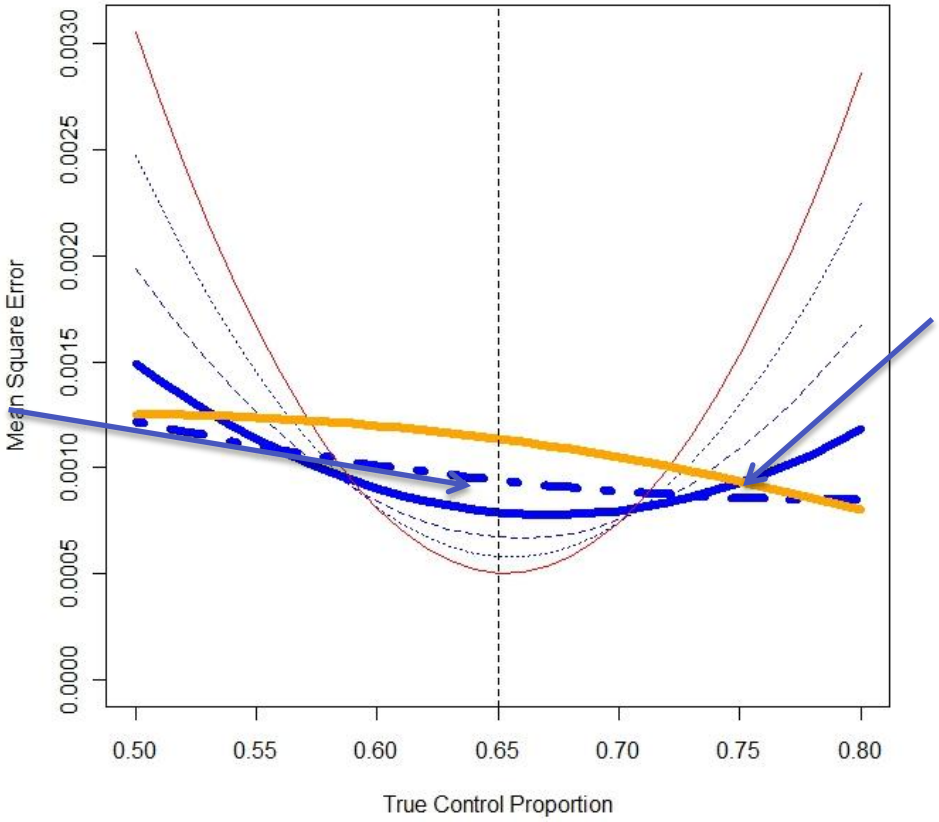
High Weight and No drift provides dramatic gains.



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

Operating characteristics for Testing

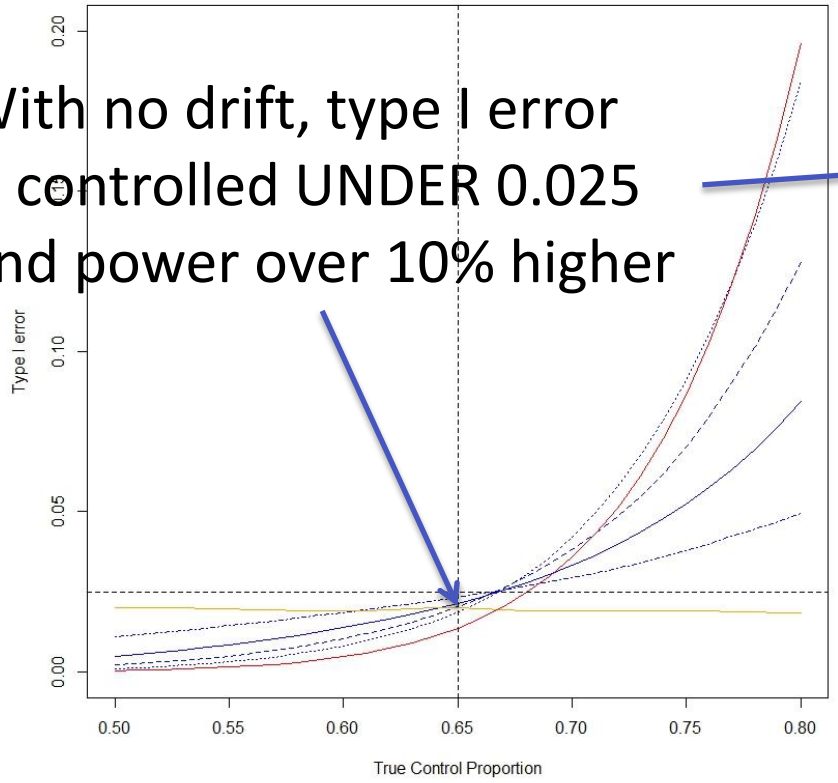
- Obtain the posterior distribution of p_C.
- Reject H_0 if $Pr(p_T > p_C) > 0.975$
  - equivalent to 95% credible interval on difference does not contain 0
  - credible interval "shifted" by historical data and weight used
- Can compute type I error and power (here we compute power for 12% gain)

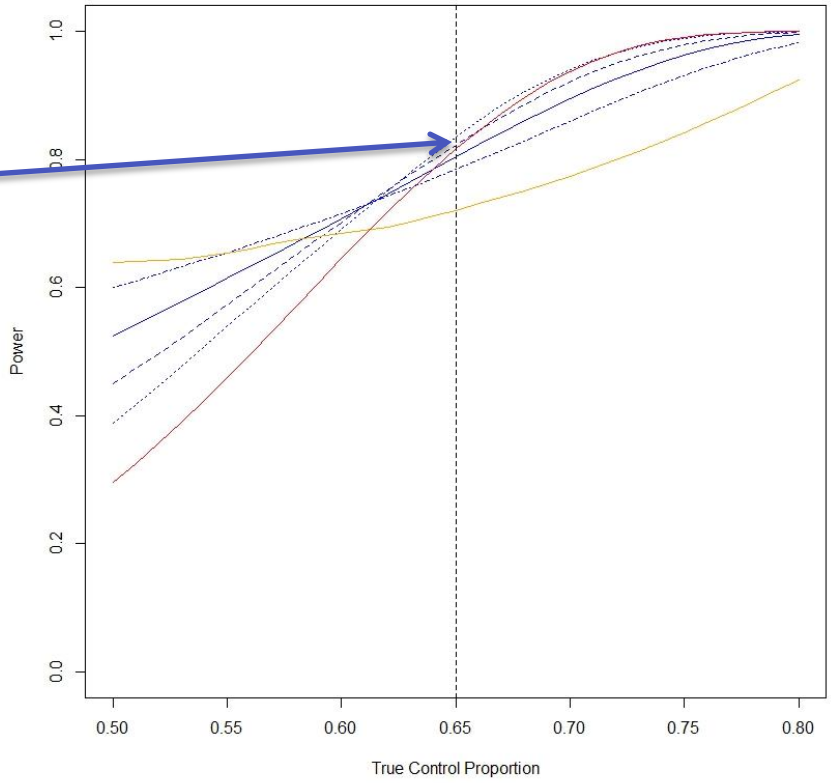
Fixed Weights (Testing)

Type I error

Power (for 0.12 gain)

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



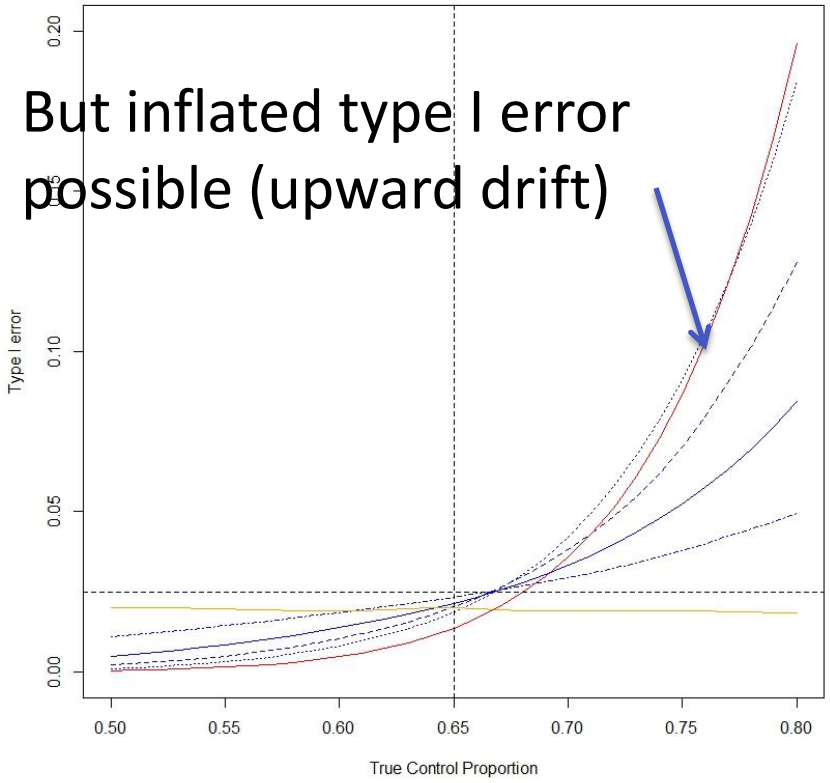


Fixed Weights

- Can use increased efficiency in multiple ways. With small amounts of drift
  - Potential 20-30% reduction in sample size
  - Can keep sample size, increase power of trial
  - Can employ unequal randomization (adaptively)

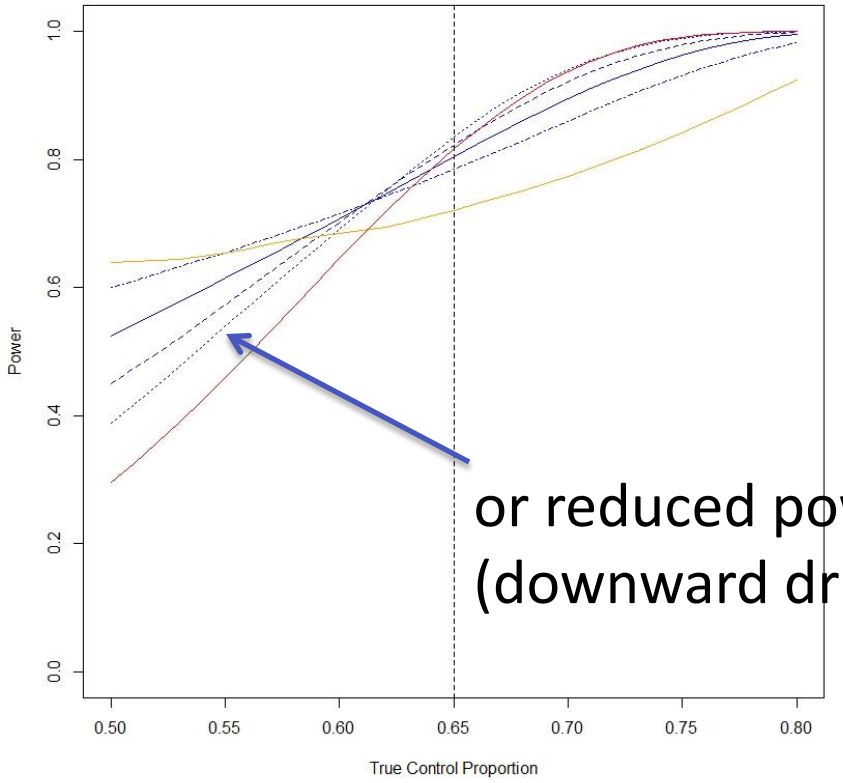
Fixed Weights (Testing)

Type I error



But inflated type I error possible (upward drift)

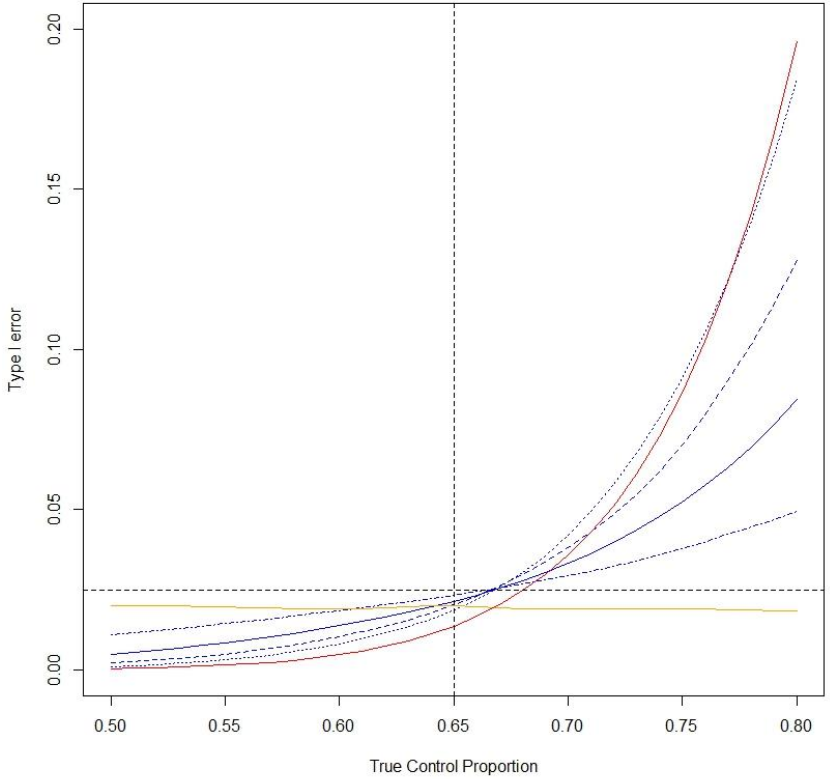
Power (for 0.12 gain)



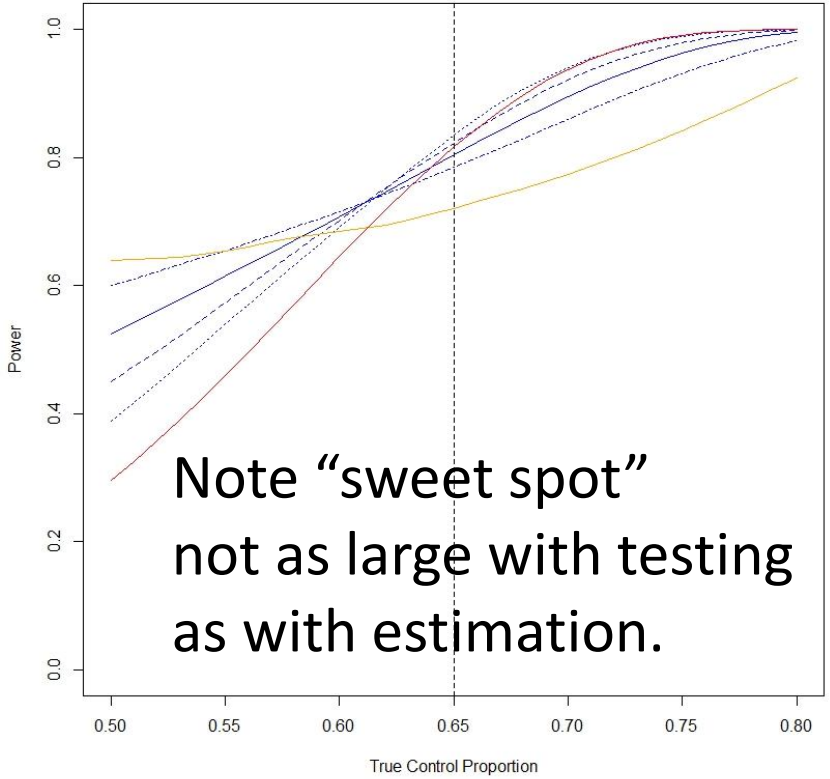
or reduced power (downward drift)

Fixed Weights (Testing)

Type I error



Power (for 0.12 gain)



Note “sweet spot”
not as large with testing
as with estimation.

Sidebar on Single Arm Trials

- Single arm trials give "infinite" weight to historical data, in that they replace the control arm with a fixed value that must be beaten
  - e.g. $H_0 : p_T=0.65$ versus $H_1 : p_T>0.65$
- No type I error inflation for null $p_T=0.65$
- but HUGE type I error inflation for the "real" question of superiority to comparator
  - e.g. null of $p_C=p_T=0.75$ in current study)

Difficult Question

- Fixed weights MIGHT pay great dividends (lower type I error, greater power OR dramatic sample size savings)
- Fixed weights MIGHT cause harm (bias, inflated type I error, reduced power)
- Can we get the best of both worlds? (not completely, but we can do better)

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

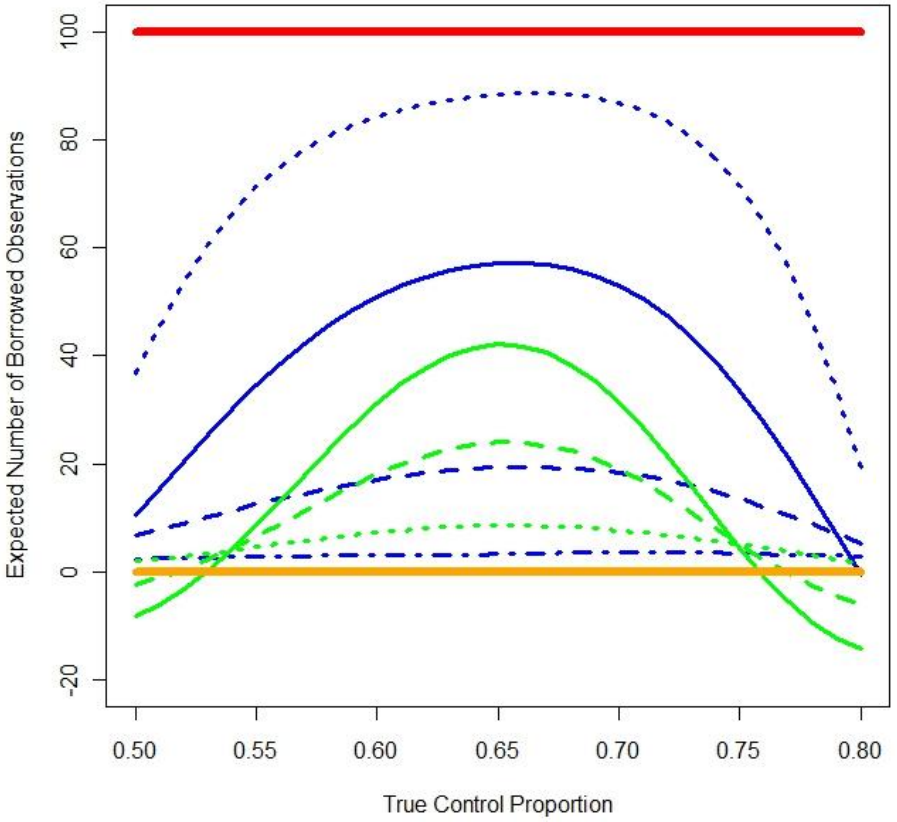
- $\text{logit}(p_C),\ldots,\text{logit}(p_H) \sim N(\mu,\tau)$
- τ measures across study variation
- We use an IGamma, here we obtained good operating characteristics.
  - other prior structures available

Hierarchical Models

- Fixed τ corresponds to specific weights.
- A prior on τ allows weight to be partially estimated from the data.
- 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)

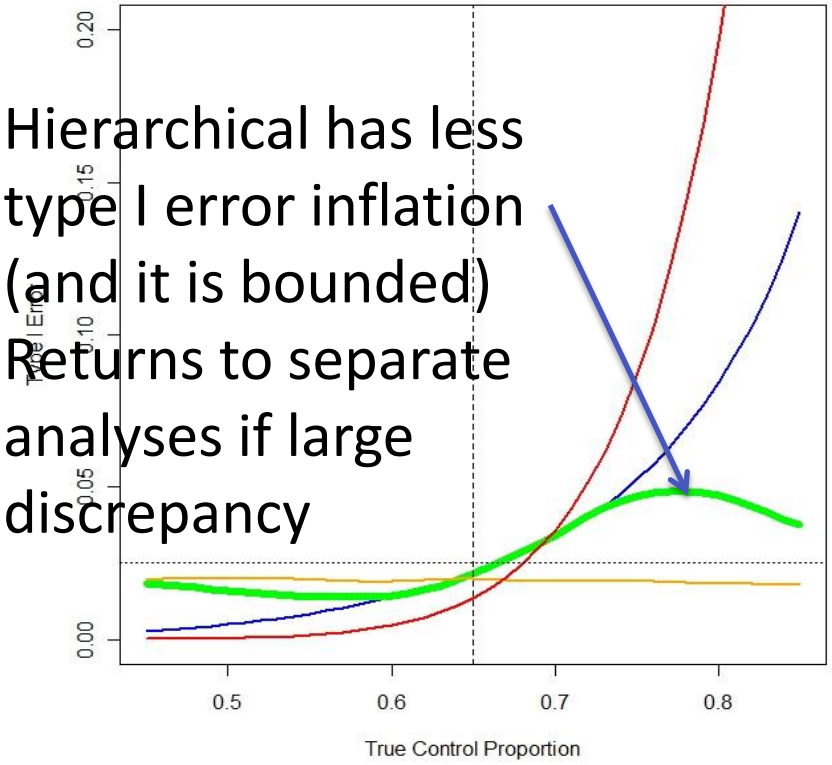
Y-axis shows expected number of borrowed subjects



Dynamic Borrowing - E[borrow] greatest for low drift

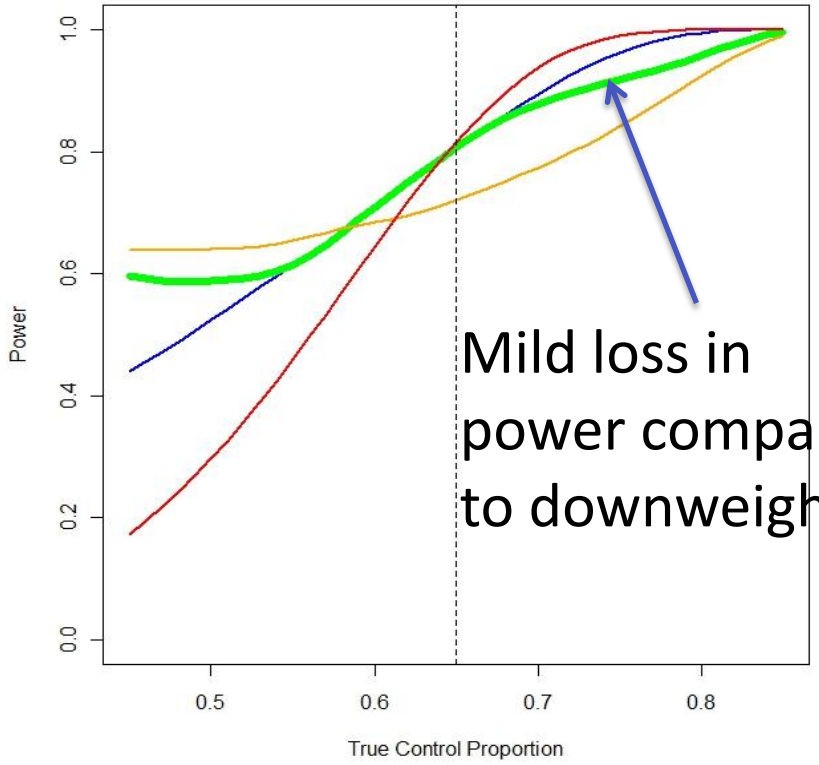
Hierarchical Models

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

Dynamic Borrowing

- Dynamic borrowing reduces and bounds risks
- However, borrowing may still help or harm the operating characteristics.
- Early phase - sponsor assessment
- Late phase – Type 1 inflation is problematic

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% gain)
  - randomized 200 (ctrl), 400 (trmt)
  - for “expected” drift, control of type I error and comparable power.

Type I error inflation

- Maximal type I error inflated by borrowing
  - dynamic is better, but still inflated.
- Type I error may be reduced
  - significantly (e.g. 0.025 to 0.017) with no drift
  - in terms of E[type 1] taken over drift dist'n.
- Minimax rules on type I error lead to unintuitive results.

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.

Surely you must pool…

- Any drift is a function of sampling variability ONLY
- Controversial?....Any rule which precludes pooling here is a questionable rule
  - if the existence of a parameter that inflates type I error precludes borrowing, you can't borrow here
  - is that a good rule?

Any difference?

- Suppose we have some idea of drift
  - pc=pc+rnorm(1,0,0.01) #pc=pc+N(0,0.01)
  - pc=pc+rnorm(1,0.01,0.01)
  - pc=pc+rnorm(1,0.1,0.05)
  - pc=pc+rnorm(1,0,0.2)
  - [assume suitably truncated to (0,1)]
- Are there some situations where you might borrow, and some where you might not?
- Not necessarily a “prior” could reflect knowledge of indication.

More realistic

- If you have information about drift, you can access whether borrowing is more or less likely to help
- Infectious diseases and antibiotics.
  - Drift is often downward due to development of resistance
- Generally, this is a medical question.
  - What do we know about the trends over time and across studies

More realistic

- Given substantive knowledge, can compute E[type I error] for any distribution on drift.
- This means, in the long run, would we make better decisions using historical data or not?
  - can't tell for any particular study, this is looking more at long run type 1 and type 2 error rates.

More realistic

- Thus, let drift~F
- Find E[type 1 error] averaging across possible values of drift
- Similarly find E[power]
- Average includes possibility of limited drift and possibility or large drift, but weighted by likelihood of occuring.
- Question…better to have?
  - 3 studies with type I error rates of 0.026,0.015,0.015
  - 3 studies with type I error rates of 0.025,0.025,0.025

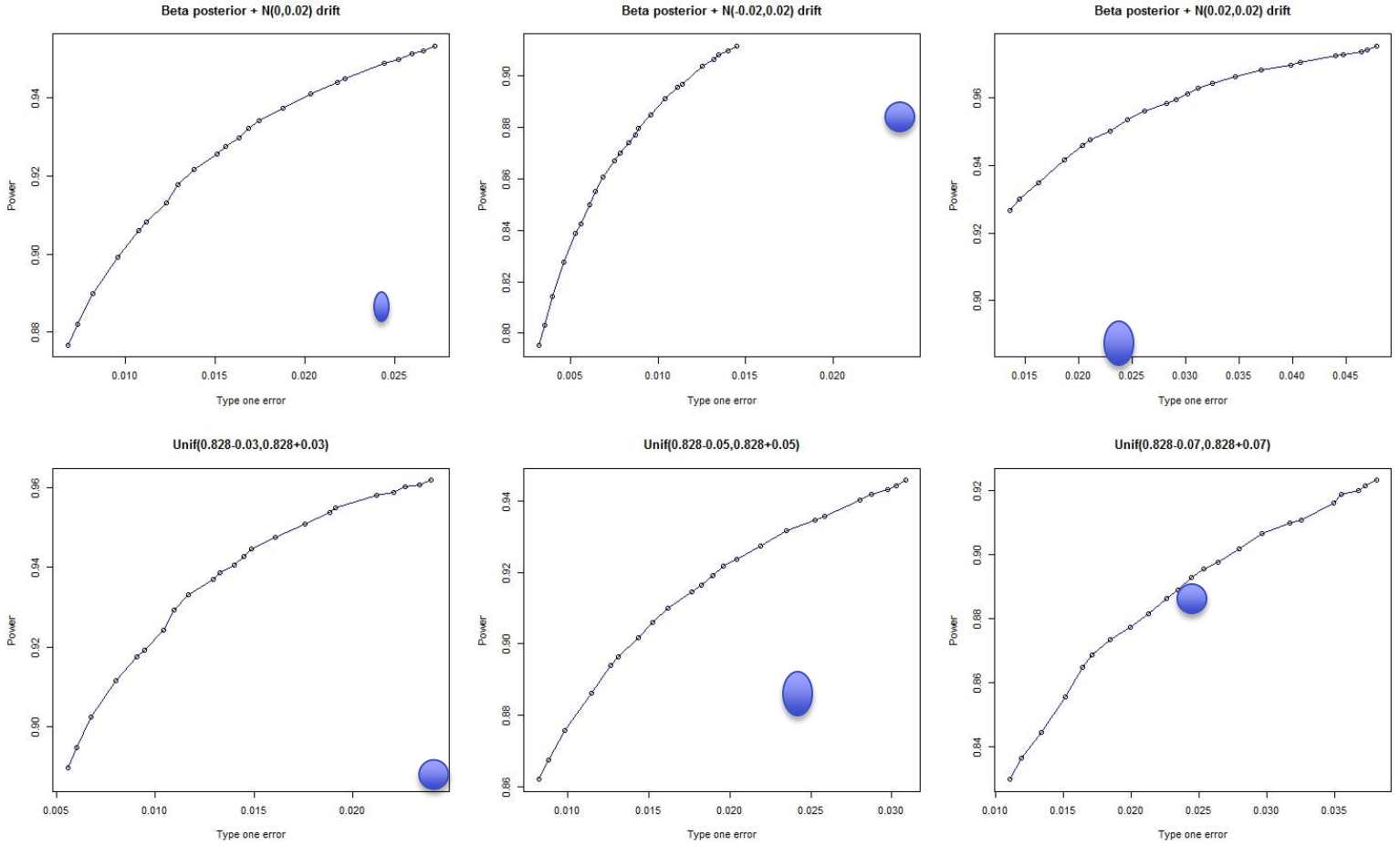
ROC Curves

- Our testing rule was
  - reject H_0 if $Pr(p_C<p_T) > 0.975$
- Can consider rules of form
  - reject H_0 if $Pr(p_C<p_T) >$ threshold
  - vary threshold
  - effectively change “nominal” alpha
  - lower thresholds lower type I error but reduce power

ROC Curves

- By varying threshold in
  - reject H_0 if $Pr(p_C<p_T) >$ threshold
- …we can construct an ROC curve, giving
  - E[type I error]
  - E[power]
- as functions of the threshold

ROC curves under various assumptions about drift



Berry Consultants
Statistical Innovation

Complicated Questions

- Complicated questions do not admit uniformly most powerful tests.
- Without a dominant test, must assess pros and cons of design.
- If we get a substantive idea of drift, can reduce long run frequency of type I errors and increase power.
  - but can't guarantee that for individual trials

What is (currently) achievable?

- You cannot
  - dominate “ignore history”
  - 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