ADAPT Summary

July 15, 2016

Work of Many

- Work presented here largely the result of discussion funded by ARLG
- co-PIs Roger Lewis and Brad Spellberg
- discussion included academics, pharmaceutical companies, FDA, BARDA, Berry Consultants

Standard Trial

- Observe collection of patients
  - HAP, VAP, cUTI, IA
  - Expect 30%, 40%, 30% in HAPVAP, cUTI, IA
  - Randomize 1:1 to ctrl:trmt
- Three potential inferential questions
  - noninferiority for sensitive pathogens (per group)
  - superiority for resistant pathogens (per group)
  - noninferiority overall (per group)

Standard Trial

- Standard analyses look at each group separately. Difficult to power
  - small sample sizes for resistant pathogens within a body site.
  - noninferiority margins are tight for sensitive pathogens within a body site
- Historically has required larger trials

Superiority on Resistant Pathogens Scenarios

- STR=sensitivity testing result
- For patients with STR=resistant to SOC and receiving SOC, expect (40%, 45%, 70%) success rates on the three body sites.
- For patients receiving novel treatment
  - consider three scenarios

Scenario	HAPVAP	cUTI	IA
Alternative	0.82	0.88	0.89
Joint Null	0.40	0.45	0.70
UTIIA	0.40	0.88	0.89

Superiority on Resistant Pathogens Power for N=300/arm

- 300 subjects/arm across HAPVAP, cUTI, IA
  - 30% HAPVAP, 40% cUTI, 30% IA
- Some proportion resistant
  - 19% HAPVAP, 20% cUTI, 25% IA
  - expect 17 resistant HAPVAP per arm, for example
- Power is low for IA, not great for HAPVAP

Power	HAPVAP	cUTI	IA
Alternative	0.762	0.913	0.419
Joint Null	0.035	0.029	0.036
UTIIA	0.035	0.913	0.419

Superiority on Resistant Pathogens Power for N=300/arm

- In standard trials, only ways to solve this are
  - increase the sample size significantly
  - enrich for resistant pathogens
    - enrichment is difficult, and involves recruiting a nonrepresentative sample of patients
    - lowers information on the other inferential questions
- We pursue innovations in trial design

Population of Antibiotics

- Suppose we have 40 drugs to test
  - 20 work in all three groups (alternative)
  - 10 don't work at all (null)
  - 10 work in UTI and IA, but not HAPVAP (UTIIA)
- If we do standard tests on the 120 drug/site combinations
  - how many drugs end up "approved" in each site?
  - how many of those are actually good?
  - how many good drugs did we miss?

Population of Antibiotics

- There are 80 good drug/site combinations
  - we get 55.2 approved on average
  - we miss a lot in IA
- There are 40 bad drug/site combinations
  - only 1.34/40 approved (3.35% type I error rate)
    - can recalibrate to 2.5%, not done here for simplicity

Separate	HAPVAP	UTI	IA
Good drugs	15.24/20.0	27.38/30.0	12.58/30.0
Bad drugs	0.70/20.0	0.28/10.0	0.36/10.0

Population of Antibiotics

- Can we do better?
- Don't want to increase the population level type I error rate.

Separate	HAPVAP	UTI	IA
Good drugs	15.24/20.0	27.38/30.0	12.58/30.0
Bad drugs	0.70/20.0	0.28/10.0	0.36/10.0

Possible Innovations

- Sharing (borrowing) information across body sites, increasing effective sample size
  - beneficial under some assumptions. Question is whether assumptions are satisfied
- Early stopping of body sites
  - if a drug is doing well/poorly in a body, stop early and use saved patients for other drugs
- Platform trial
  - Multiple drugs in comparison at once. Save control arm subjects as well as other efficiencies
- In combination, these can save 45-60% of subjects compared to a standard trial

Borrowing

- With only 1 group in a trial (and good trial practices), results are generally "unbiased"
  - equal chance the trial results are higher or lower than the truth.
- With multiple groups, each individual group may be better estimated using the information from all groups.
  - Statistical work from 1950s (Stein, others)
  - General intuition (if first 2 of 3 groups are successful, you may be more optimistic about the 3rd)
  - Properties of random error (next slide)

Adding noise creates excessively large highs and lows

- Often illustrated with sports analogies
  - if you have lots of players, the best performers are typically "lucky AND good". If you want to estimate their real attributes accurately, you have to estimate away the luck. This applies to poor performance as well.
- Statistically, imagine three groups
  - True success rates 0.70, 0.75, 0.80
  - Observe 20 observations from each
    - the average low is 66%, the average high is 84%
    - you can get better estimates by "shrinking" (66%,84%) together, closer to the true (70%, 80%)

Simple Model

- We have also been discussing a more elaborate model.
  - simplified here to isolate borrowing
- Let γ_b be the success rate on SOC in each body site, and let $\gamma_b+\theta_b$ be the success rate on the novel treatment.
- The treatment effects are the three θ_b terms
  - hierarchical borrowing
  - $\theta_0,\theta_1,\theta_2 \sim N(\mu,\tau)$ with μ and τ having prior distributions.

Hierarchical Model

- $\theta_0,\theta_1,\theta_2 \sim N(\mu,\tau)$ with μ and τ having prior distributions.
- Creates dynamic borrowing through τ
  - borrow more when data are consistent with common treatment effects, less when one the body sites looks different than the others.
- Extremes
  - forcing τ large estimates the group separately
  - forcing τ small emulates pooling
  - we avoid both these extremes

In practice what does this do?

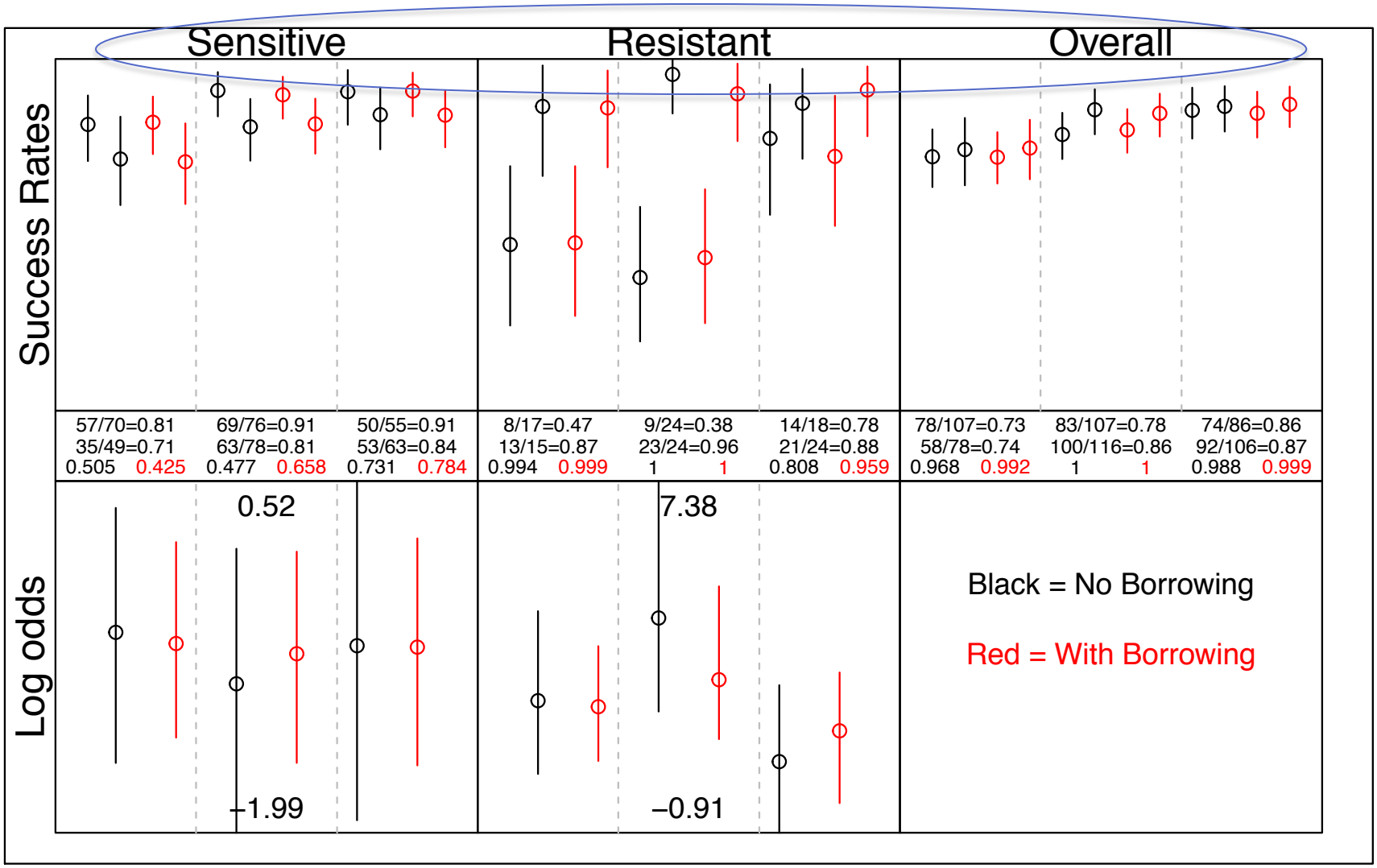
- Point estimates of treatment effect pulled closer together across body sites
  - amount dictated by model estimates of "true variation across body site" versus "random noise"
    - overlapping standard confidence intervals across groups suggest similar true effects which differ by random noise. Borrow more.
    - divergent standard confidence intervals across groups suggest real differences. Borrow less, pull the point estimates less.
- Confidence intervals shortened as all groups contribute to the estimates.
  - shorter confidence intervals means higher effective sample sizes
- Long statistical history (1950s) of advantageous performance for such estimators.

Individual Trials

- We ran 100 simulated trials and analyzed the data from each using
  - separate analyses
  - the model with borrowing
- Looked through those trials for examples of
  - where the analyses produced the same decision.
  - where a decision changed from “not success” to “success” because of the borrowing.
  - where two body sites had great results and one had poor results

Example 1

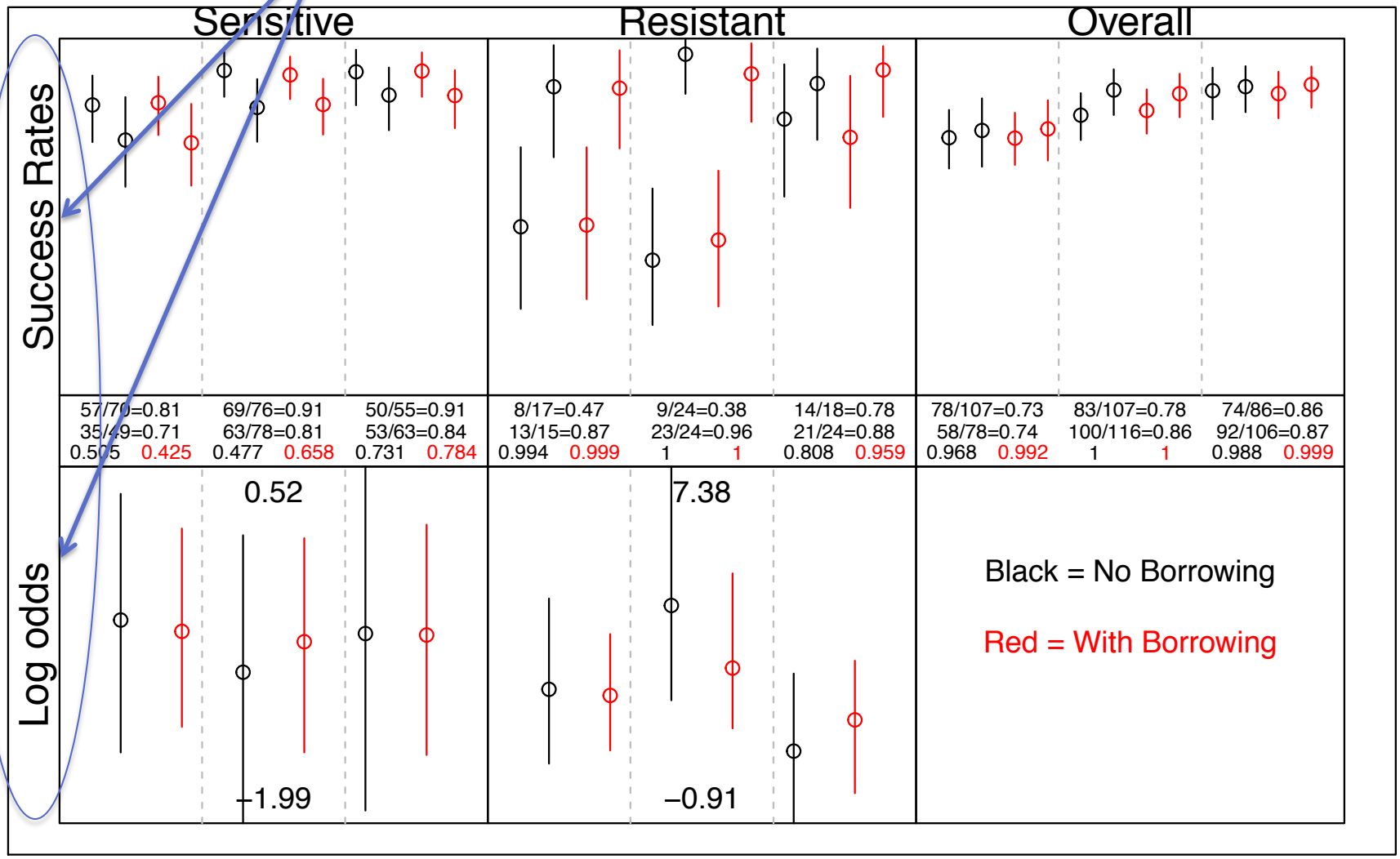
Main columns show the populations of interest…sensitives, resistants, and the overall population



Example 1

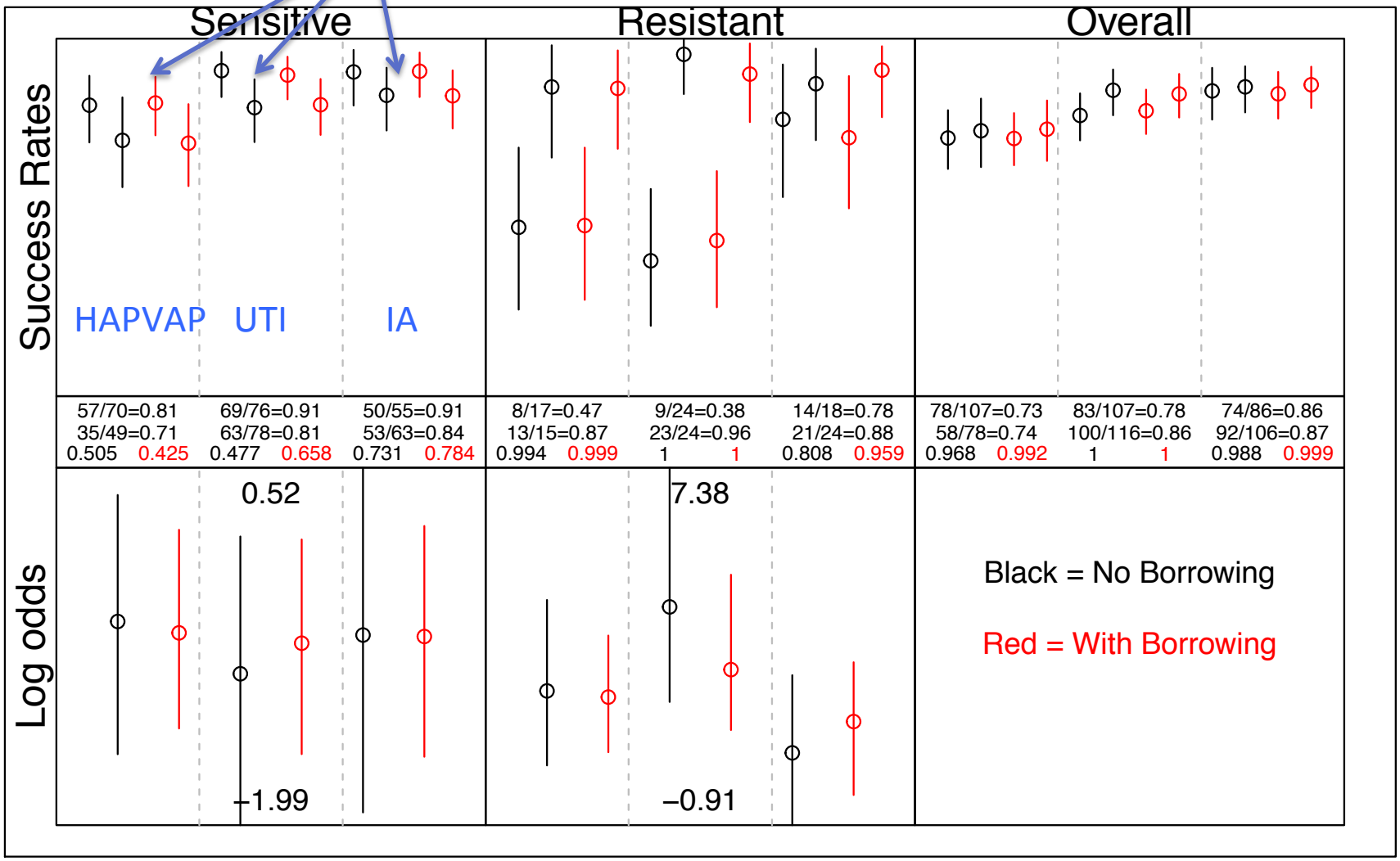
Top row shows the estimated success rates

Bottom row shows the log odds comparing treatment to control.



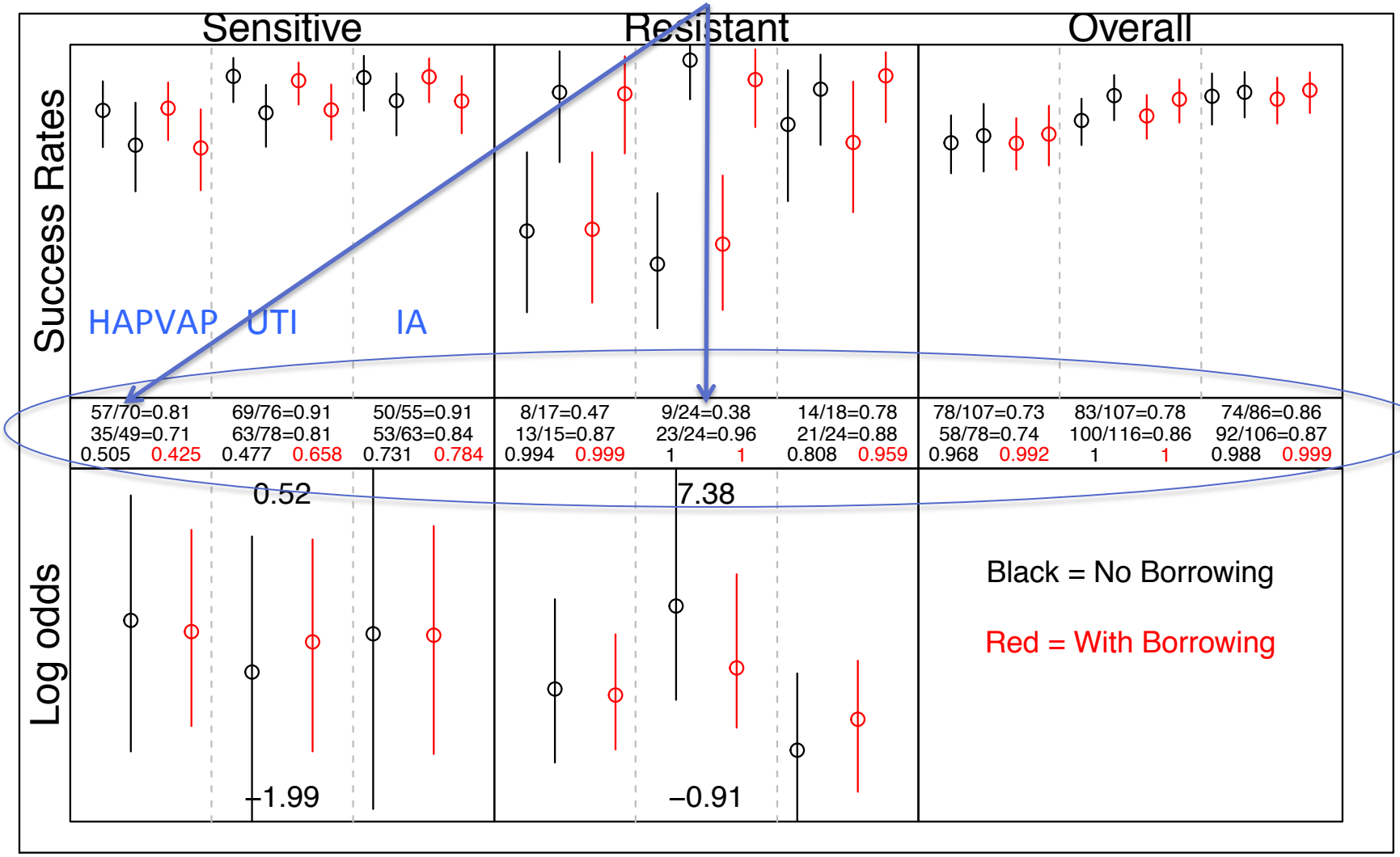
Example 1

Each main box contains 3 regions separated by dashed lines. These indicate the three body sites (HAPVAP, UTI, IA)



Example 1

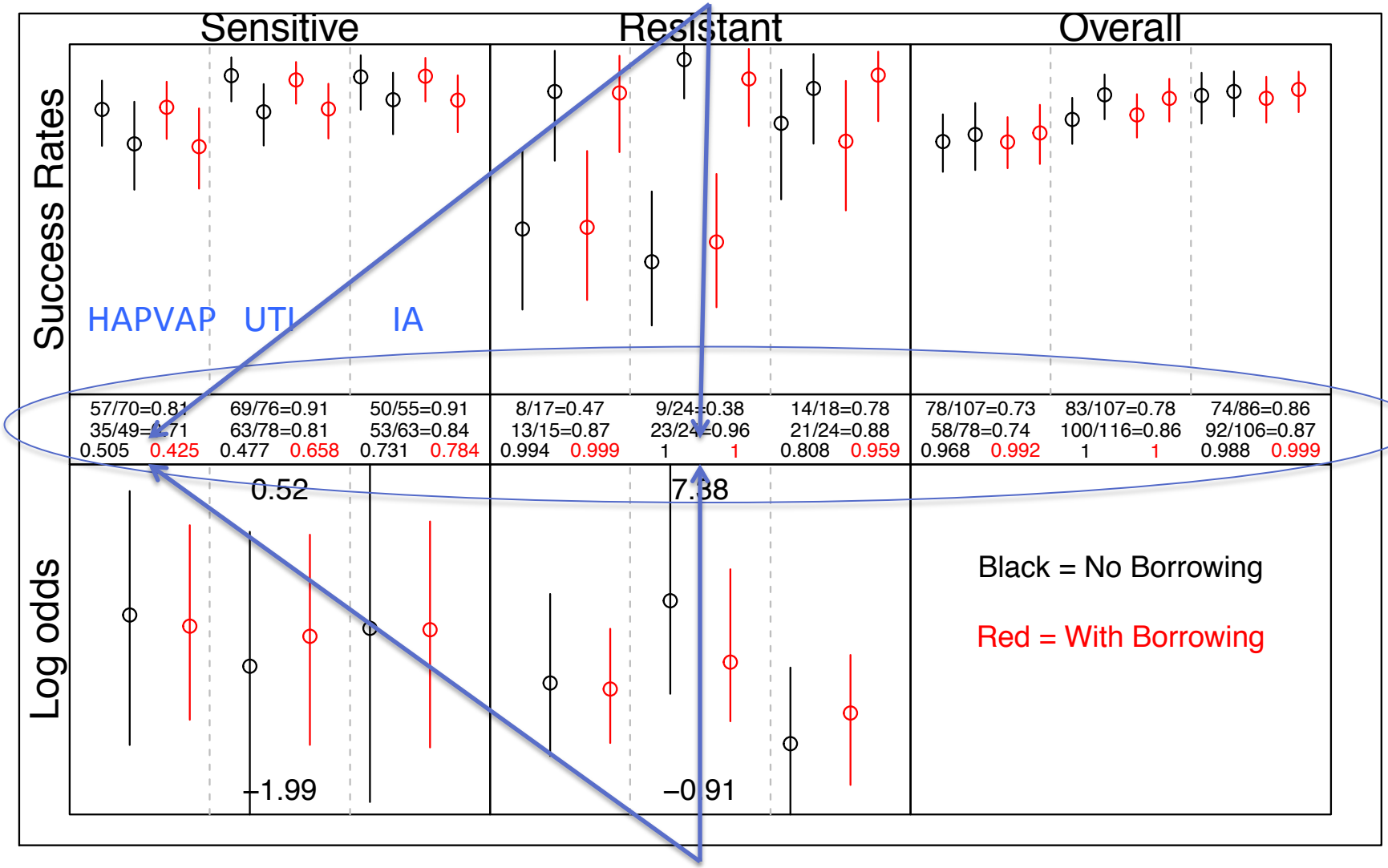
The data appears in the middle (lots to show!). The top is control, middle is treatment.
For example here Resistant UTI had 9/24 on control, and 23/24 on treatment. Good!
But HAPVAP sensitives are more marginal (57/70 control, 35/49 treatment)



Example 1

The bottom row in the middle shows the analysis of interest for each population. Either Pr(noninferiority) for sensitives or overall, or Pr(superiority) for resistants.

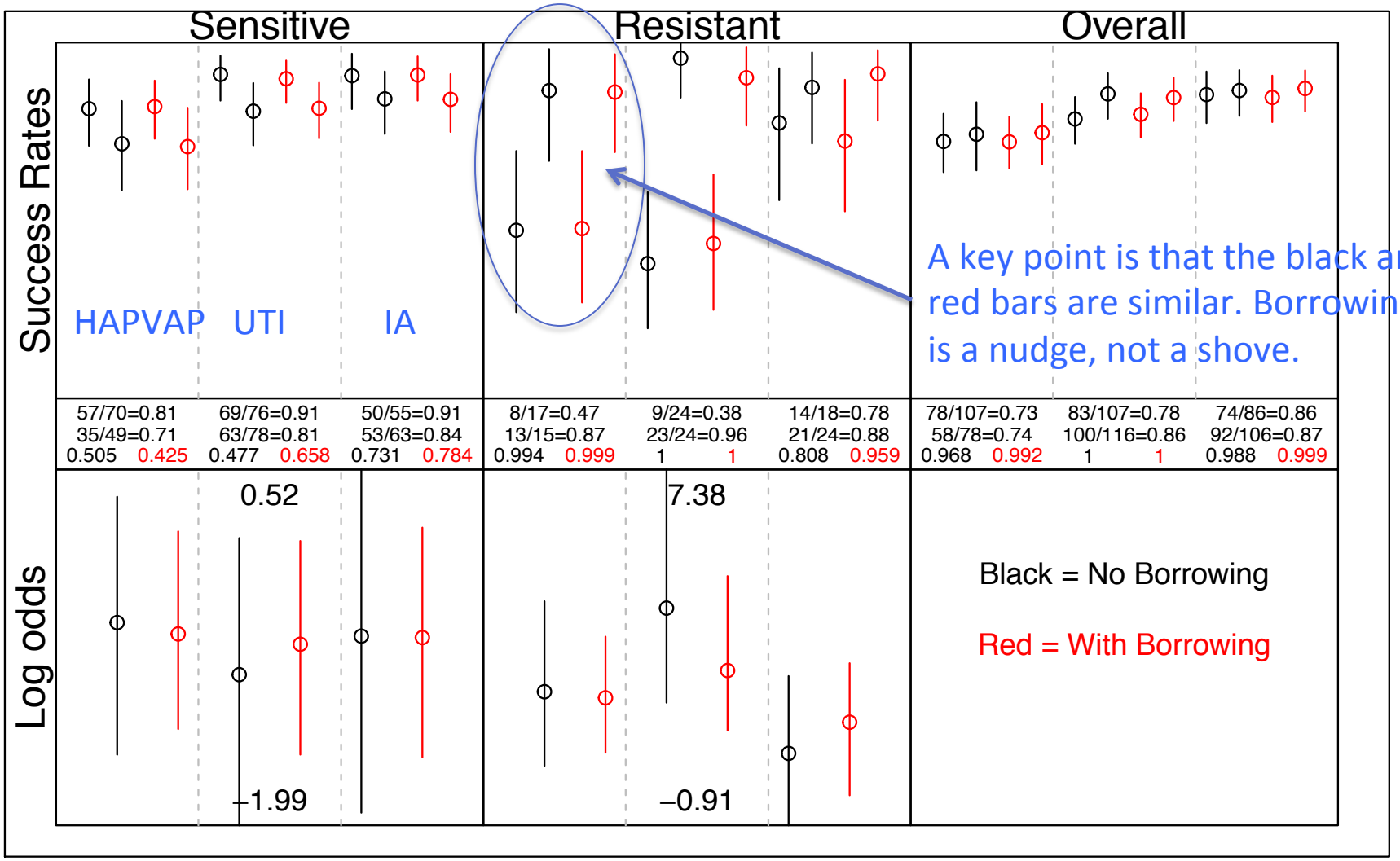
Black shows without borrowing, Red with borrowing.



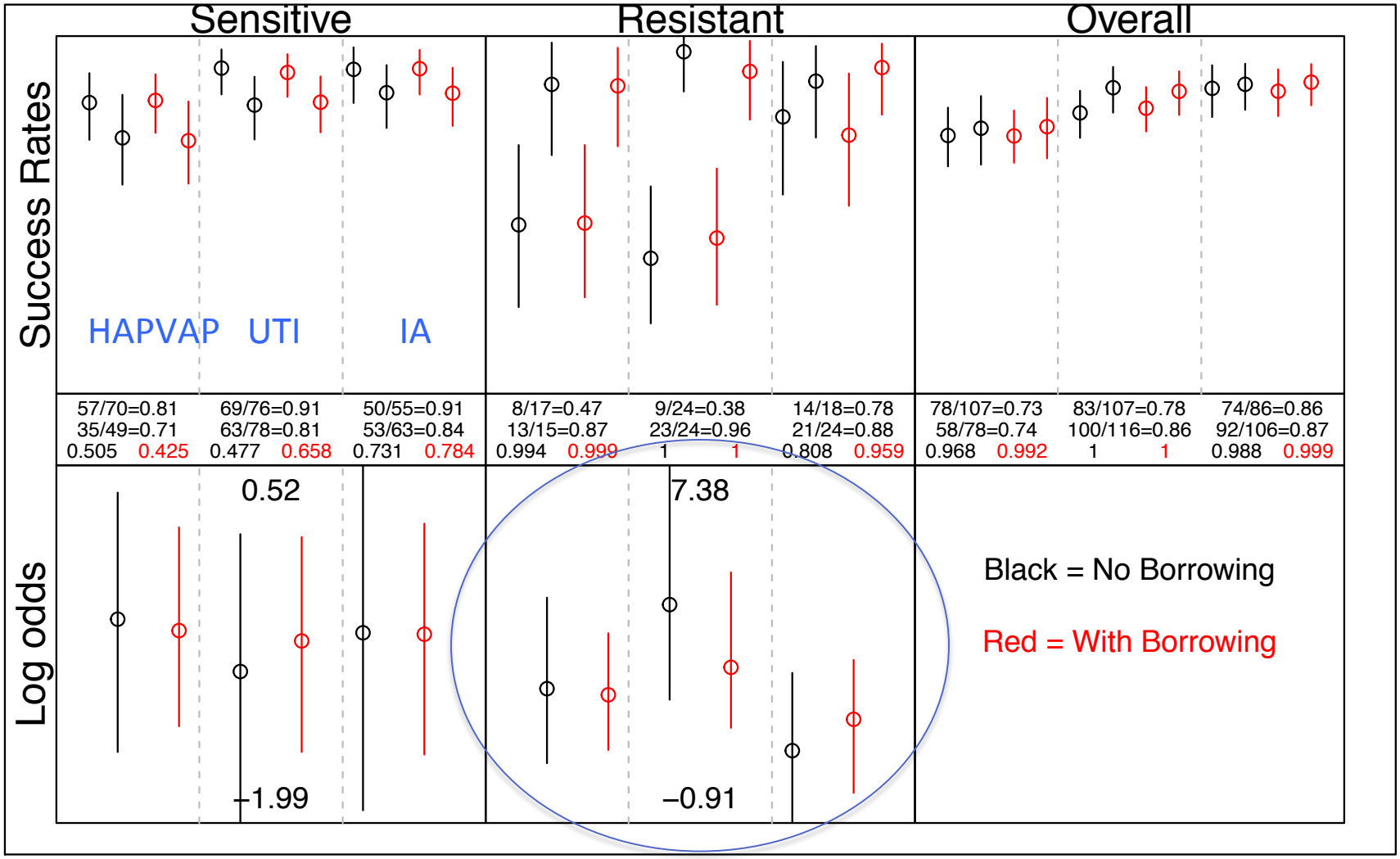
For resistant in UTI, probability almost 100%. For sensitives in HAPVAP, results not strong

Example 1

Within rates for each body site and population, you see four bars showing point estimates and confidence intervals. The first two in black show control and treatment estimates without borrowing, the next two in red show control and treatment estimates with borrowing.



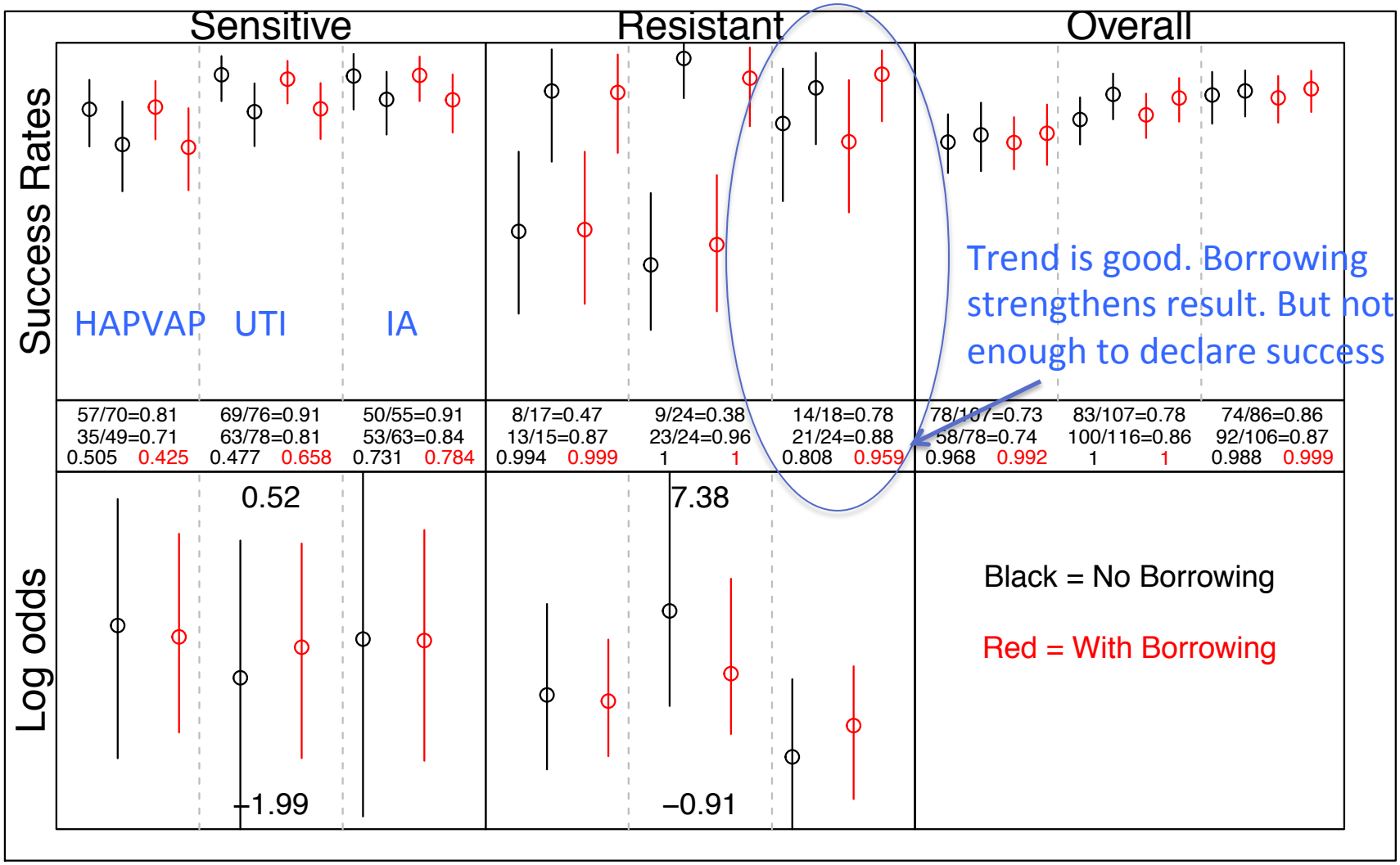
Example 1



The model borrows on the log odds scale. Within sensitives and resistants, you can see the point estimates get closer and the confidence intervals shorten (this is main advantage). Here the model discounts the 96% success rate in resistant UTI as more likely luck than real.

Example 1

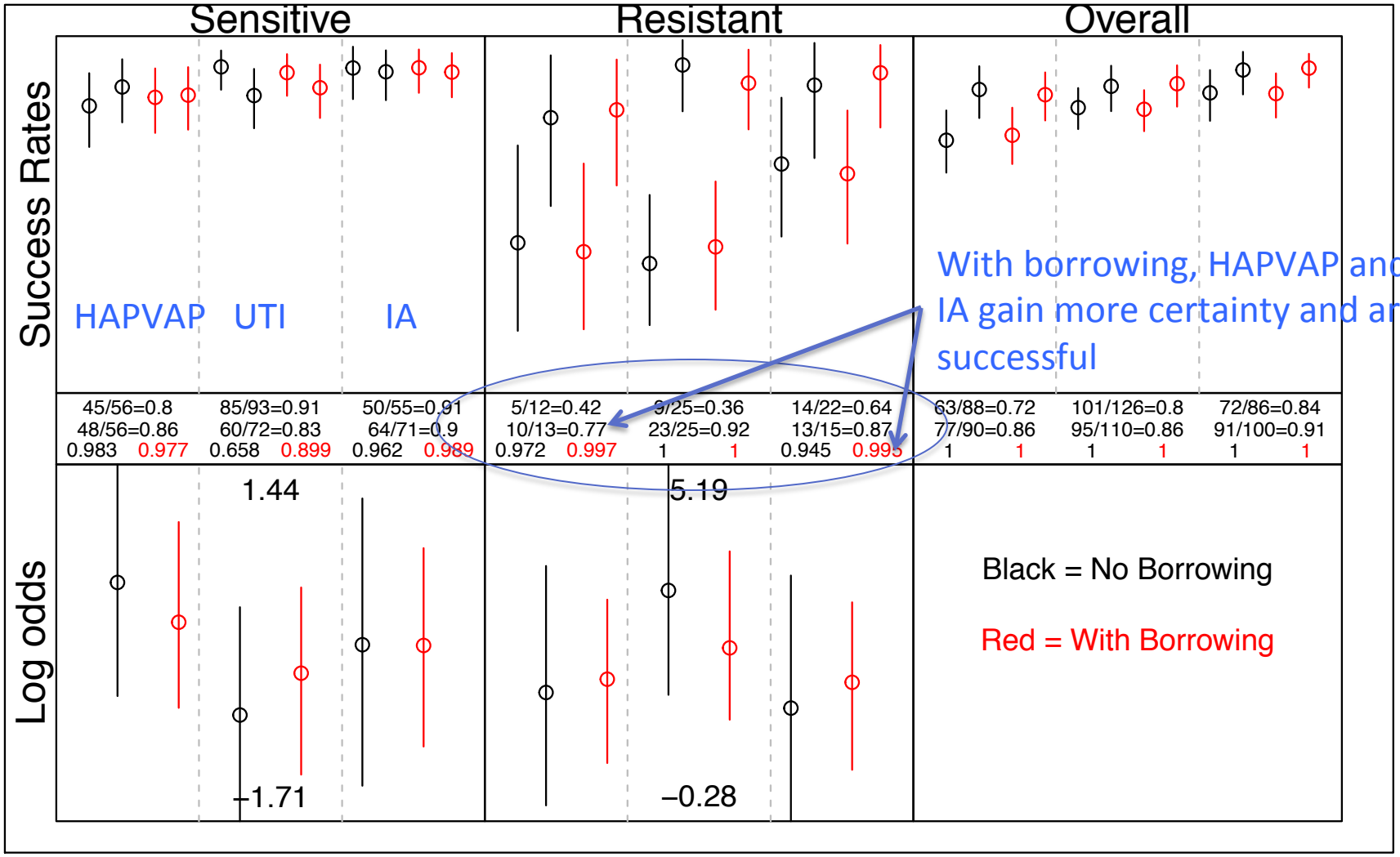
The main change HERE from borrowing is the Resistant IA superiority. But neither analysis is successful. Probability is 80% without borrowing and 96% with borrowing (not 97.5%). Data is decently in the right direction in IA and the other groups are strong in resistants.



Example 2

Example 2

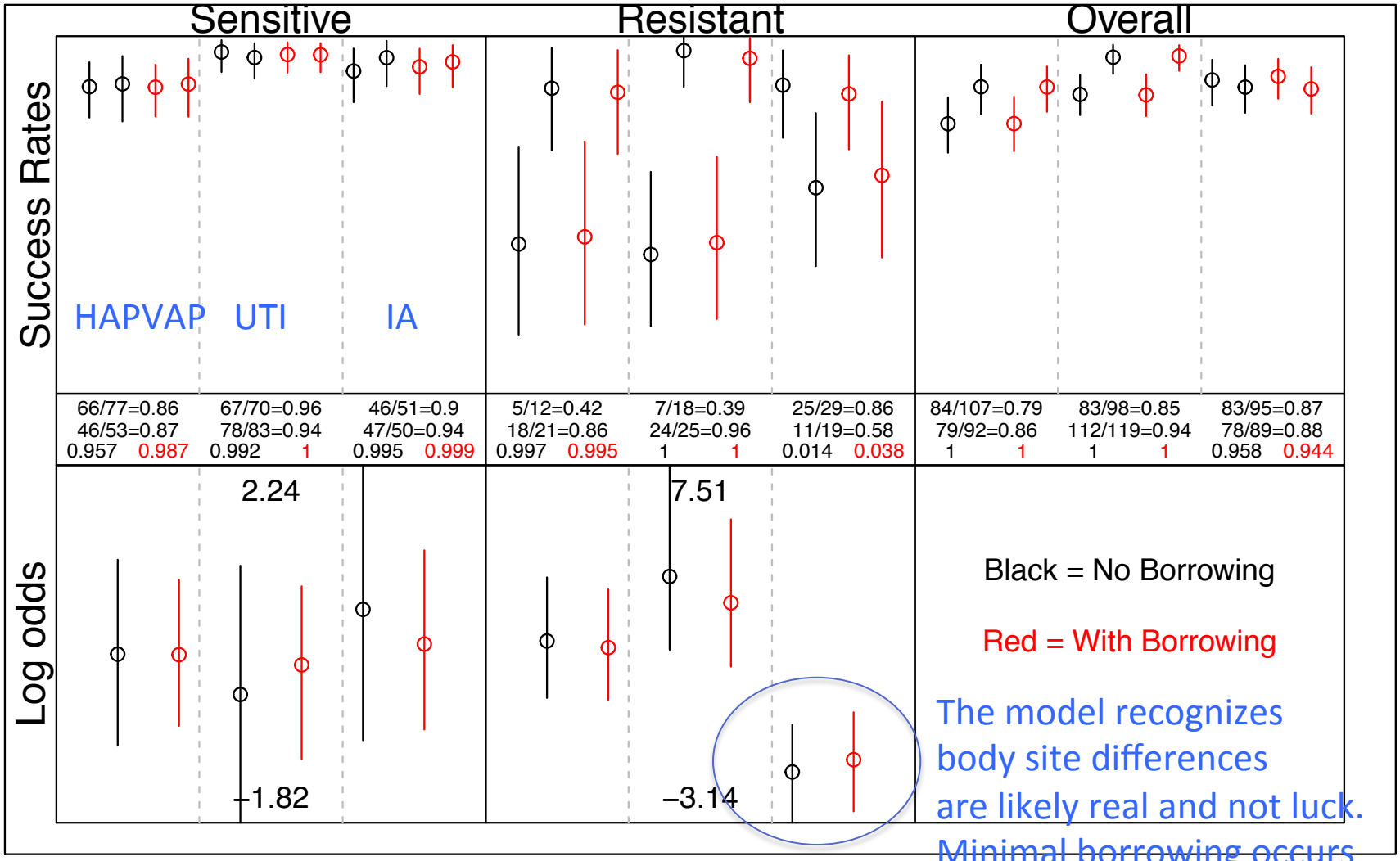
This example has a similar flavor to example 1, but the data are stronger.
Without borrowing, pvalues are about 0.028, <0.001, and 0.055 in each site for resistants, would only result in a UTI success without borrowing.



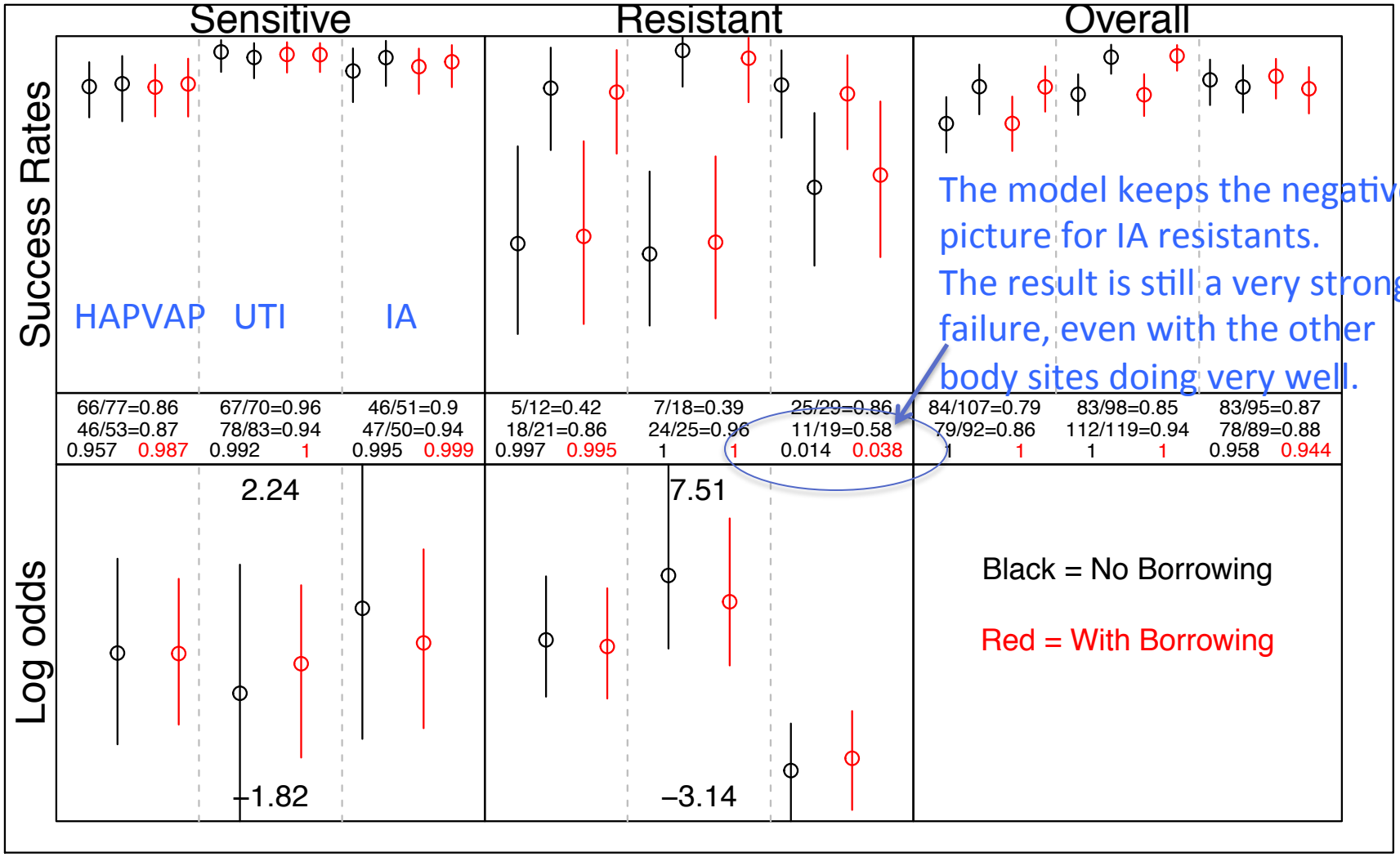
Example 3

Example 3

Here the resistant IA data is bad (it goes in the wrong direction). But the HAPVAP and UTI data is quite good…..how much does the model "move up" the bad IA results?



Example 3


Sensitive
Resistant
Overall
Success Rates
HAPVAP
UTI
IA
66/77=0.86
46/53=0.87
0.957 0.987
67/70=0.96
78/83=0.94
0.992 1
46/51=0.9
47/50=0.94
0.995 0.999
5/12=0.42
18/21=0.86
0.997 0.995
7/18=0.39
24/25=0.96
1 1
25/29=0.86
11/19=0.58
0.014 0.038
84/107=0.79
79/92=0.86
1 1
83/98=0.85
112/119=0.94
1 1
83/95=0.87
78/89=0.88
0.958 0.944
Log odds
2.24
−1.82
7.51
−3.14
The model keeps the negative picture for IA resistants.
The result is still a very strong failure, even with the other body sites doing very well.
Black = No Borrowing
Red = With Borrowing

Summary of example trials

- Borrowing not intended to dramatically alter results. Provides shift away from extreme point estimates, more certainty.
- "Dynamic" borrowing recognizes when one body site is very different than the rest.
- Generally the "extra" successes from borrowing (the extra power) results from modest shifts, not large ones.
  - can look at individual trials and see likely changes in advance to obtain comfort

Operating Characteristics

- Generally sharing of information is quite advantageous with common (identical not needed) effects, good or bad.
- Can be problematic in mixed scenarios
- Note trial can be tailored to avoid mixed scenarios to some degree.
  - drugs which appear not to adequately reach pathogens in the lung in earlier clinical work would be tested in only cUTI and IA, for example.
  - "mixed" only refers to the body sites actually being investigated in trial.

N=300/arm with borrowing

- "Borrowing" is a continuum. We chose a particular prior for "moderate borrowing"

Power (Sep/Borrow)	HAPVAP	cUTI	IA
Alternative	0.762 / 0.925	0.913 / 0.979	0.419 / 0.753
Joint Null	0.035 / 0.016	0.029 / 0.021	0.036 / 0.018
UTIIA	0.035 / 0.155	0.913 / 0.898	0.419 / 0.543

1) Note increase in power in alternative for all groups is substantial (76% to 93% in HAPVAP, 42% to 75% in IA)
2) Type 1 error is REDUCED substantially in all groups in the joint null
3) In the mixed scenario power is mildly increased for IA, mildly decreased for cUTI
4) The cost is the 15.5% error rate for HAPVAP when treatment is effective in UTI and IA but not HAPVAP
5) Essentially odds of making correct decision improved in 7 cells, close in 1, worse in 1

How to weigh risks and benefits?

- In our hypothetical population of 40 drugs we had 120 drug/site combinations
  - depends strongly on proportion of drugs in each category of underlying truth.
  - better chance of right answer for 100
    - better power, lower type I error
  - similar chance of right answer in 10
  - worse chance in 10
    - inflated type I error rate

N per arm=300 (with and without borrowing)

Power (N=300) Borrowing/Separate	HAPVAP	cUTI	IA
Alternative	0.762/0.925	0.913/0.979	0.419/0.753
Joint Null	0.035/0.016	0.029/0.021	0.036/0.018
UTIIA	0.035/0.155	0.913/0.898	0.419/0.543

Expected proportion of population treated correctly after experiment (e.g. success for effective groups and nonsuccess for ineffective groups).

	Separate	Borrowing
Alternative	72.0%	89.5%
Joint Null	96.7%	98.1%
UTIIA	78.0%	77.6%

Long run behavior
Several years….

- Return to our population of drugs
  - 20 work in all three groups (alternative)
  - 10 work in UTI and IA, but not HAPVAP (UTIIA)
  - 10 don't work at all (null)
- Success rates (N=300/arm) per group

Separate	HAPVAP	UTI	IA
Good drugs	15.24/20.0	27.38/30.0	12.58/30.0
Bad drugs	0.70/20.0	0.28/10.0	0.36/10.0
Borrowing	**HAPVAP**	**UTI**	**IA**
Good drugs	18.50/20.0	28.56/30.0	20.48/30.0
Bad drugs	1.70/20.0	0.22/10.0	0.36/10.0

93.86/120 correct decisions

105.26/120 correct decisions

Long run behavior
Several years….

- For every 1 “extra” incorrect HAPVAP approvals from borrowing
  - 3 extra good drugs on shelf for HAPVAP
  - 1 extra good drugs on shelf for UTI
  - 8 extra good drugs on shelf for IA
- Trade 1 type 1 error for 12 more correct approvals
  - can optimize level of borrowing for such a criterion.

Long run behavior
Several years….

- Suppose you wanted NO extra type I errors.
- Can recalibrate borrowing
- Another solution is to raise the bar for success.
  - currently Pr(trmt>ctrl) > 0.975
  - recalibrate to Pr(trmt>ctrl) > 0.986

Power (N=300/arm) Sep/0.975/0.986	HAPVAP	cUTI	IA
Alternative	0.762/0.925/0.882	0.913/0.979/0.957	0.419/0.753/0.670
Joint Null	0.035/0.016/0.008	0.029/0.021/0.010	0.036/0.018/0.011
UTIIA	0.035/0.155/0.108	0.913/0.898/0.846	0.419/0.543/0.438

Long run behavior
Several years….

- For every 40 drugs
  - original threshold had 1 extra type I error and 12 extra true successes
  - new threshold has 0 extra type I errors and 7.8 extra true successes.

Power (N=300/arm) Sep/0.975/0.986	HAPVAP	cUTI	IA
Alternative	0.762/0.925/0.882	0.913/0.979/0.957	0.419/0.753/0.670
Joint Null	0.035/0.016/0.008	0.029/0.021/0.010	0.036/0.018/0.011
UTIIA	0.035/0.155/0.108	0.913/0.898/0.846	0.419/0.543/0.438

Long run behavior
Several years….

- For every 40 drugs
  - original threshold had 1 extra type I error and 12 extra true successes
    - to recalibrate separate trials to do this, change alpha=0.05 and increase sample size to 425 (increase of 42% from N=300 per arm)
  - new threshold has 0 extra type I errors and 7.8 extra true successes.
    - to recalibrate separate trials to do this, no change in alpha and increase sample size to 425-450 as well

Early Stopping of Body Sites

- Suppose a drug performs poorly in HAPVAP but is still promising in UTI and IA.
  - stop enrolling patients in HAPVAP on that drug, but continue other body sites
- Similarly, body sites that perform well can be stopped early for success
  - with multiple looks must account for the type I error implications

Early Stopping example (superiority for resistant only)

- Perform interim analyses at N=300,400,500 enrolled subjects
- Fit hierarchical borrowing model to all the observed data
  - stop body site for futility if Pr(trmt > ctrl) < 0.5 (essentially if observed data is worse on the treatment arm)
  - stop body site for success if Pr(trmt > ctrl) > 0.995
- If body site doesn't stop, at end of trial drug is successful in body site if Pr(trmt>ctrl) > 0.975 (later we consider other thresholds)

Early Stopping
final threshold 0.975

Power (N=300/Arm) Sep/Bor975/ Bor975early	HAPVAP	cUTI	IA
Alternative	0.721/0.903/0.934	0.914/0.978/0.979	0.423/0.747/0.812
Joint Null	0.038/0.015/0.019	0.030/0.019/0.022	0.035/0.018/0.021
UTIIA	0.038/0.158/0.168	0.914/0.899/0.904	0.423/0.545/0.615

Modest increases in power/type I error reflecting additional looks to win

Sample sizes reduced 20% in "joint" scenarios, 7.5% in mixed scenario

Average Sample Size Sep/Bor975early	HAPVAP	cUTI	IA	Total
Alternative	180 / 143	240 / 156	180 / 182	600 / 481
Joint Null	180 / 141	240 / 183	180 / 141	600 / 465
UTIIA	180 / 189	240 / 179	180 / 188	600 / 556

Population behavior

- Return to population of 40 drugs
  - 20 work in all three body sites (alternative)
  - 10 don't work anywhere (joint null)
  - 10 UTIIA (work in UTI and IA, not in HAPVAP)
  - in total, of 120 body site/drug combinations, have 80 that work, 40 that don't
- Standard trial, on average
  - 1.41/40 type I errors and 54.53/80 true successes
  - N=24,000 subjects
- Early Stopping with threshold 0.975
  - 2.30/40 type I errors and 69.69/80 true successes
  - N=19,830 subjects (17% reduction, could test 8 more drugs)

Calibration

- Calibrating the early stopping design to achieve the same type I error rate in this population of drugs requires adjusting threshold to around 0.990

N=300 per arm per drug	Type 1 errors (out of 40 possible)	True successes (out of 80 possible)	Total Sample Size
Standard Design	1.41	54.53	24000
Bor975early	2.30	69.69	19830
Bor986early	1.67	65.67	19830
Bor990early	1.40	62.56	19830

Sample size savings

- What sample size would be needed in a standard trial to achieve the same power (true success rate) increase?

	Type 1 errors (out of 40 possible)	True successes (out of 80 possible)	Total Sample Size
Standard Design N=300 per arm	1.41	54.53	24000
Standard Design N=425 per arm	1.27	61.80	34000
Bor986early Max N=300 per arm	1.67	65.67	19830
Bor990early Max N=300 per arm	1.40	62.56	19830

Platform advantages

- Previous slides talked focused on separate trials for each drug, for both standard design and borrowing with early stopping.
  - both had half control, half treatment
  - for 24,000 subjects, 12,000 were control
- In a platform design, multiple drugs are tested at once.
  - novel treatments move in and out of the trial
  - control arm enrolls the entire time.

Platform advantages

- Suppose always enrolling control and 4 drugs

Control					
1	9	13	16	17	and so on
2	6	10	14	and so on	
3	5	8	11	15	18	and so on
4	7	12	and so on		

- Instead of enrolling 1:1 for 40 drugs, enroll 1:1:1:1:1 at any given time.
  - Net result, instead of 24000 subjects (12000 control and 12000 treatment), still need 12000 treatment, but only 3000 control subjects (shared control)

Platform advantages

- Sample size savings immediate from shared control (15000 subjects from 24000).
- Early stopping saved 17% of the sample size for compared to borrowing alone, can still borrow and adaptively stop drugs in platform
  - saving 17% of the 12000 treatment subjects results in needing 10000 treatment subjects. Results in 13000 subject trial.

Platform advantages

- These are “back of the envelope” calculations
- More details/rigor in Saville and Berry in slightly different context (Clinical Trials 2016, “Efficiencies of platform clinical trials: A vision of the future” currently online ahead of print)

Platform caveats

- Often treatments are compared to "concurrent controls" (controls enrolled during time treatment in platform)
  - This induces correlation among the treatments. If the controls are randomly low/high in one stretch of time, affects multiple drugs at once.
  - Modeling the entire control sequence may mitigate this concern (must account for "drift")
- Likely will encounter periods of time where full benefit of trial can't be achieved, e.g. no drugs available to fill an empty slot for some length of time.
- Obviously logistically complex (but examples exist)
- What is the control arm/s? Can it be changed?

Summary

- Adding early stopping to borrowing can reduce sample sizes
  - standard design requires 400-425 per arm
  - borrowing alone reduced sample sizes to 300 per arm.
  - early stopping as well reduces that to 230-275 per arm.
- A platform trial structure produces further advantages
  - sharing control information
  - utilizing subject savings to accelerate investigation of future drugs.
  - 13000 subjects over 40 drugs average 325/drug (not arm)