PHARMALEX

CASE STUDY
Issues In Regulatory Acceptability Of Bayesian Historical Data Borrowing Methods

Lauren J. Frazee, PharmaLex US
Angel Lu, GSK***
Bradley P. Carlin, PharmaLex US

October 25, 2023

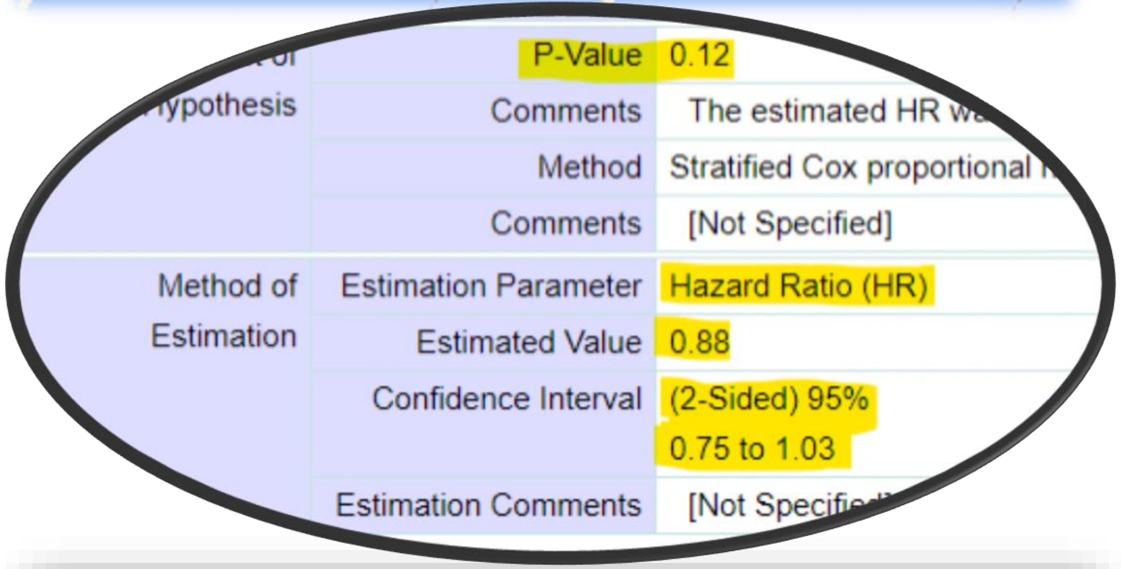
Disclaimers

- (From previous slide: ***This work was done outside of GSK.)
- The information provided during this presentation does not constitute legal advice. PharmaLex, and its parent Cencora, strongly encourage the audience to review available information related to the topics discussed during the presentation and to rely on their own experience and expertise in making decisions related thereto. Further, the contents of this presentation are owned by PharmaLex and reproduction of the slides used in today's presentation is not permitted without consent of PharmaLex.
- *Conflict of interest statement:* Bradley P. Carlin is a member of the local organizing committee for BAYES2023.

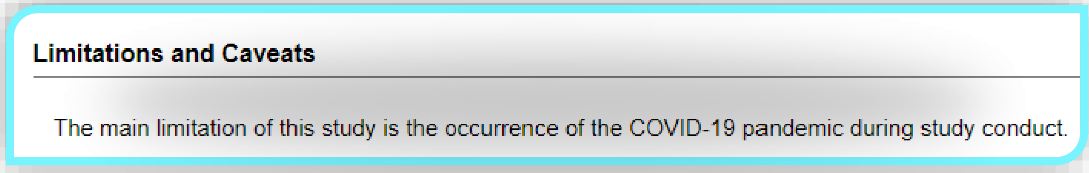
Relevant Background (1)

- Our client wanted to submit an investigational new drug application to the US FDA for the design and planned analyses of a prospective Phase III trial for a drug designed to reduce rates of myocardial infarction in subjects with a particular genetic profile and recent Acute Coronary Syndrome (ACS).

- This client's previous Phase III clinical trial for this drug showed a near-significant treatment effect, but ultimately **failed its primary endpoint.**
  - COVID-related setbacks may have contributed to this failure



	P-Value	0.12
ypothesis	Comments	The estimated HR wa
	Method	Stratified Cox proportional
	Comments	[Not Specified]
Method of Estimation	Estimation Parameter	Hazard Ratio (HR)
	Estimated Value	0.88
	Confidence Interval	(2-Sided) 95% 0.75 to 1.03
	Estimation Comments	[Not Specifie



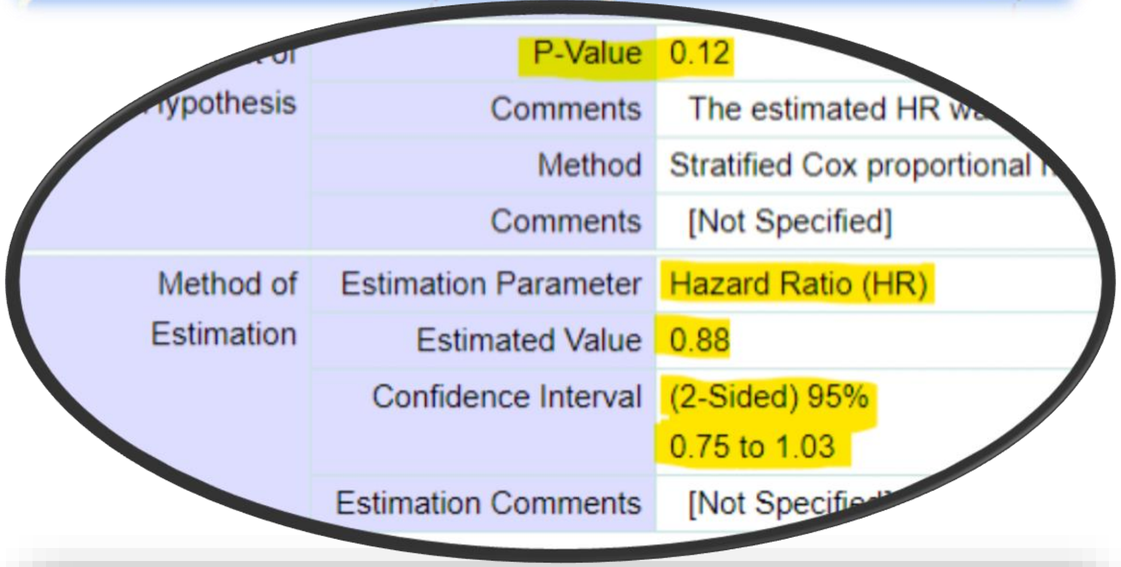
Limitations and Caveats

The main limitation of this study is the occurrence of the COVID-19 pandemic during study conduct.

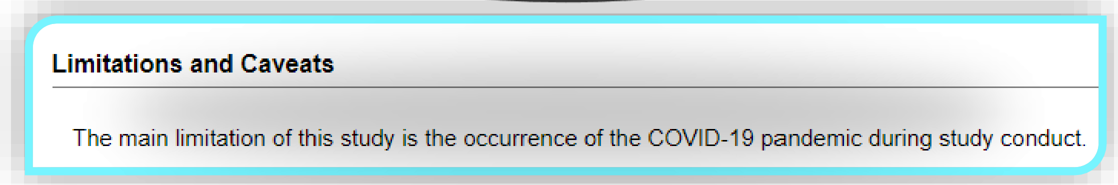
Unmodified images publicly available at: https://clinicaltrials.gov

Relevant Background (2)

- The client wished to try again in a subpopulation shown to be promising in Study 1 (some “data snooping”…)
- To increase power while reducing new trial costs, we proposed an adaptive design featuring partial borrowing from the first study’s efficacy data.
- Initially, FDA regulators expressed willingness to consider a “development program-wide” assessment of trial power and Type I error
  - i.e., we would control α at .025 one-sided across both studies, an unconditional Type I error assessment



ypothesis	P-Value	0.12
	Comments	The estimated HR wa
	Method	Stratified Cox proportional h
	Comments	[Not Specified]
Method of Estimation	Estimation Parameter	Hazard Ratio (HR)
	Estimated Value	0.88
	Confidence Interval	(2-Sided) 95% 0.75 to 1.03
	Estimation Comments	[Not Specifie



Limitations and Caveats

The main limitation of this study is the occurrence of the COVID-19 pandemic during study conduct.

Unmodified images publicly available at: https://clinicaltrials.gov

Bayesian Meta-Analytic Survival Model

- To combine information from the two studies, we propose fitting a Weibull survival model to each study separately, obtaining a normal approximation to the treatment effect δ in the first (1) and second (2) trials, respectively, i.e.

$$p(\delta|t_1) \approx N\left(\hat{\eta}_1, \widehat{P_1}\right) \qquad \text{and} \qquad p(\delta|t_2) \approx N\left(\hat{\eta}_2, \widehat{P_2}\right)$$

- The Bayesian Central Limit Theorem[1] ensures that these approximations will be accurate using mean and precision estimates from the MCMC algorithm used to fit the model.
- We assume Study 2 always enters at full weight, while Study 1 has weight $w \in (0,1)$
- Using Bayes Rule, the two normal distributions for δ can be merged into a single normal,

$$p(\delta|t_1, t_2) \approx N(\eta_c, P_c)$$

where $P_c = w\,\widehat{P_1} + \widehat{P_2}$ is the total combined precision

and $\eta_c = (w\,\widehat{P_1}\,\widehat{\eta_1} + \widehat{P_2}\,\widehat{\eta_2})/P_c$ is a weighted average of the two study-specific means.

Choosing *w* using unconditional Type I error (1)

- Using Stoffer’s method (an extension of Fisher’s meta-analytic method for combining p-values), we found the overall, 2-sided p-values that would result from using potential weights of the new study (*w*) along with the observed p-value from the previous trial in combination with potential 1-sided p-values that might be observed in the new study:

New trial 1-sided p-values: w	**0.01**	**0.025**	**0.05**	**0.075**	**0.10**
0	0.020	0.050	0.100	0.150	0.200
0.1	0.011	0.029	0.061	0.095	0.130
0.2	0.006	0.017	0.037	0.060	0.085
0.3	0.004	0.010	0.024	0.039	0.056
0.4	0.002	0.007	0.016	0.026	0.038
0.5	0.002	0.005	0.011	0.019	0.027
0.6	0.001	0.004	0.008	0.014	0.020
0.8	0.001	0.003	0.006	0.009	0.013
1.0	0.001	0.002	0.004	0.007	0.010

Choosing *w* using unconditional Type I error (2)

- If the new trial achieves a one-sided p-value of 0.025, then a weight *w* as low as **0.1** will still deliver an overall two-sided p-value less than 0.05.
- Even if the new trial delivers a one-sided p-value of 0.10, a weight *w* as low as **0.4** would work

New trial 1-sided p-values: ***w***	**0.01**	**0.025**	**0.05**	**0.075**	**0.10**
0	0.020	0.050	0.100	0.150	0.200
0.1	0.011	**0.029**	0.061	0.095	0.130
0.2	0.006	0.017	0.037	0.060	0.085
0.3	0.004	0.010	0.024	0.039	0.056
0.4	0.002	0.007	0.016	0.026	**0.038**
0.5	0.002	0.005	0.011	0.019	0.027
0.6	0.001	0.004	0.008	0.014	0.020
0.8	0.001	0.003	0.006	0.009	0.013
1.0	0.001	0.002	0.004	0.007	0.010

Adaptive Trial Design: Stopping Rules

- To create a stopping rule for the Bayesian meta-analysis, we use a Bayesian Z-score to compute the tail area

$$p = P(\delta > 0 \,|t_1, t_2) = P\left(Z_c > -\eta_c\sqrt{P_c} \,\middle|\, t_1, t_2\right) = \Phi\left(\eta_c\sqrt{P_c}\right),$$

where Φ is the standard normal distribution function.

- Our adaptive trial stops and declares efficacy at the first (n = 120 events) interim look if $p > q_1$, where q_1 is selected so that this happens just 0.1% of the time when δ = 0.
  - In (just) this interim calculation, *w* is set equal to 0, meaning the new trial data alone must deliver the strong significance required for early stopping.

- At the final (n = 200 events) look, we stop and declare final efficacy if $p > q_2$, where q_2 is selected so that this happens just an additional 2.4% of the time when δ = 0.
  - Thus the overall program-wide alpha level is controlled at 2.5%, one-sided (hence 5% two-sided).

- Our design estimates p through repeated sampling of J = 1000 datasets from the null, taking q_1 and q_2 as the appropriate empirical quantiles of the p distribution.

Adaptive Trial Design: Power at *w* = 0.1 (1)

- We simulated J = 1000 ‘alternative’ (i.e., not null) datasets to find the impact of each of 3 potential effect sizes in the new trial, along with the weight (*w*) of the old trial, applied during the final look only.

- Below, we show overall probabilities of success, or power, in each alternative scenario; the numbers in the parentheses correspond to the probabilities of stopping at the interim and final looks, respectively:

HR w	0.7	0.75	0.8	Stopping probability thresholds (q_1, q_2)
0.0	0.75 (0.22, 0.53)	0.66 (0.17, 0.49)	0.56 (0.13, 0.43)	(0.997, 0.971)
0.05	0.80 (0.22, 0.58)	0.73 (0.17, 0.56)	0.63 (0.13, 0.50)	(0.997, 0.966)
0.1	0.82 (0.22, 0.60)	0.75 (0.17, 0.58)	0.65 (0.13, 0.52)	(0.997, 0.966)
0.15	0.85 (0.22, 0.63)	0.78 (0.17, 0.61)	0.69 (0.13, 0.56)	(0.997, 0.964)
0.2	0.87 (0.22, 0.65)	0.81 (0.17, 0.64)	0.73 (0.13, 0.60)	(0.997, 0.962)

Adaptive Trial Design: Power at *w* = 0.1 (2)

- Here we have defined statistical ‘significance’ as exceeding a 0.997 posterior probability of effectiveness at the interim analysis (120 events), or a 0.9666 probability of effectiveness at the final analysis (200 events).
  - So we have 82% power to detect an HR of 0.7, with 22% of those stops coming early
  - These probabilities correspond to interim and final look Type I error-spending probabilities of 0.001 and 0.024, for a total of 0.025 one-sided.

HR w	0.7	0.75	0.8	Stopping probability thresholds (q_1, q_2)
0.0	0.75 (0.22, 0.53)	0.66 (0.17, 0.49)	0.56 (0.13, 0.43)	(0.997, 0.971)
0.05	0.80 (0.22, 0.58)	0.73 (0.17, 0.56)	0.63 (0.13, 0.50)	(0.997, 0.966)
0.1	**0.82 (0.22, 0.60)**	**0.75 (0.17, 0.58)**	0.65 (0.13, 0.52)	(0.997, 0.966)
0.15	0.85 (0.22, 0.63)	0.78 (0.17, 0.61)	0.69 (0.13, 0.56)	(0.997, 0.964)
0.2	0.87 (0.22, 0.65)	0.81 (0.17, 0.64)	0.73 (0.13, 0.60)	(0.997, 0.962)

Final Accepted Trial: *w* = 0.025

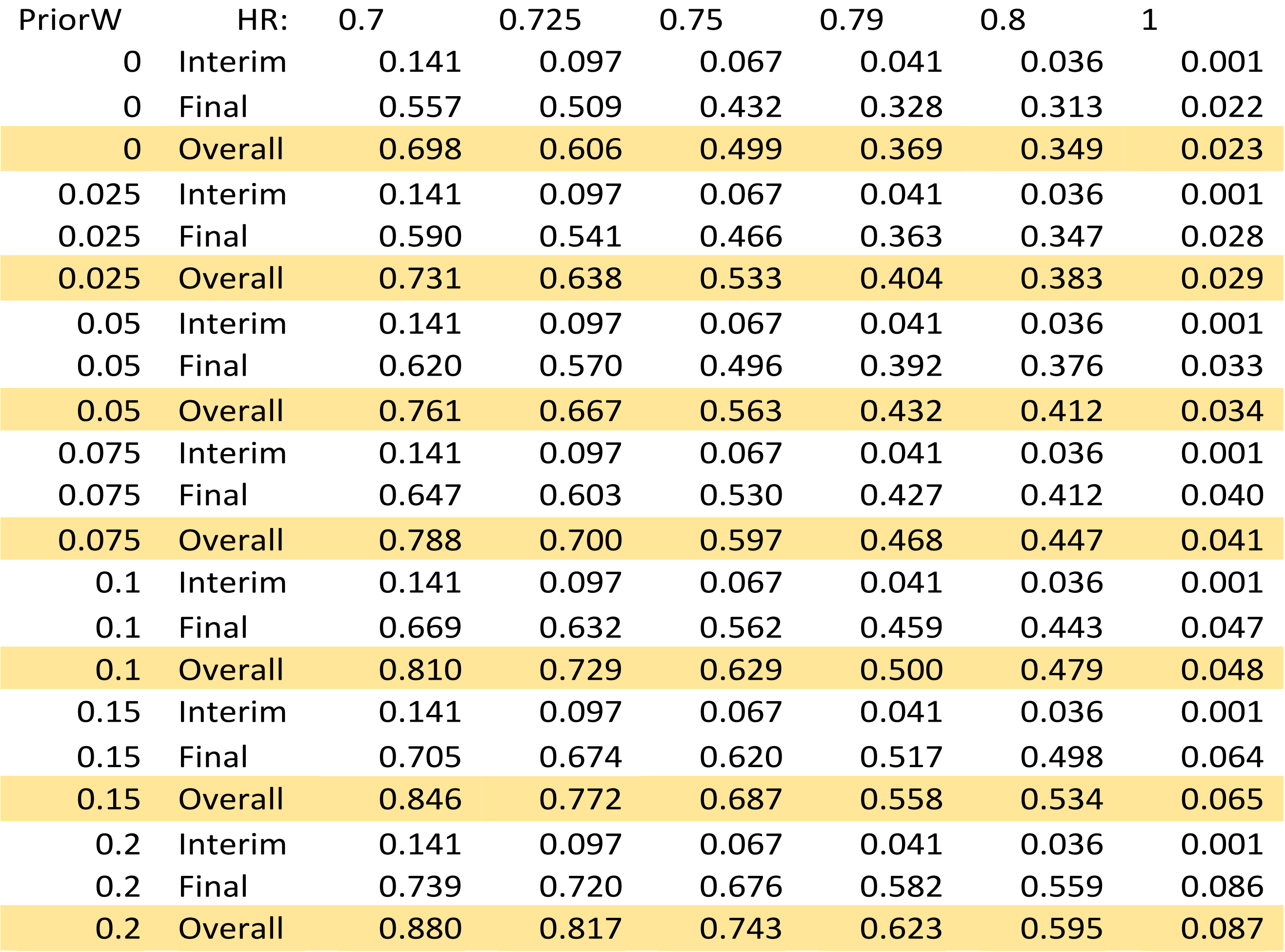
- Unfortunately, our proposal to down-weight the previous trial data to just 10% of that of the prospective trial in the meta-analysis was still met with significant pushback.

- Despite their earlier position, ultimately the program director insisted on limiting *conditional* Type I error inflation to just 10% (i.e., permitting an increase to just 0.0275 from 0.025).
  - This decision seemed to be driven by the following results, as requested by the agency, harkening back to Stouffer's method:

w	New trial (only) p-value (two-sided)	New trial (only) p-value (one-sided)	HR required in new trial	p-value from Stouffer's weighted Z test statistic
0.025	**0.0573**	**0.0287**	**0.764**	**0.05**
0.05	0.0653	0.0327	0.770	0.05
0.075	0.0740	0.0370	0.776	0.05
0.1	0.0834	0.0417	0.782	0.05

Conditional Power Table for Bayesian MA Design

- First column is *w* (proportion of information borrowed)
- Second column tells whether these are interim look, final look, or overall (interim + final) powers
- The final column (HR=1) gives one-sided conditional Type I error
- Remaining columns give the conditional powers for the various true HR values
- Thus, the 2-sided conditional Type I error when *w = 0.025* is 2*0.029 = **0.058**.



PriorW	HR:	0.7	0.725	0.75	0.79	0.8	1
0	Interim	0.141	0.097	0.067	0.041	0.036	0.001
0	Final	0.557	0.509	0.432	0.328	0.313	0.022
0	Overall	0.698	0.606	0.499	0.369	0.349	0.023
0.025	Interim	0.141	0.097	0.067	0.041	0.036	0.001
0.025	Final	0.590	0.541	0.466	0.363	0.347	0.028
0.025	Overall	0.731	0.638	0.533	0.404	0.383	0.029
0.05	Interim	0.141	0.097	0.067	0.041	0.036	0.001
0.05	Final	0.620	0.570	0.496	0.392	0.376	0.033
0.05	Overall	0.761	0.667	0.563	0.432	0.412	0.034
0.075	Interim	0.141	0.097	0.067	0.041	0.036	0.001
0.075	Final	0.647	0.603	0.530	0.427	0.412	0.040
0.075	Overall	0.788	0.700	0.597	0.468	0.447	0.041
0.1	Interim	0.141	0.097	0.067	0.041	0.036	0.001
0.1	Final	0.669	0.632	0.562	0.459	0.443	0.047
0.1	Overall	0.810	0.729	0.629	0.500	0.479	0.048
0.15	Interim	0.141	0.097	0.067	0.041	0.036	0.001
0.15	Final	0.705	0.674	0.620	0.517	0.498	0.064
0.15	Overall	0.846	0.772	0.687	0.558	0.534	0.065
0.2	Interim	0.141	0.097	0.067	0.041	0.036	0.001
0.2	Final	0.739	0.720	0.676	0.582	0.559	0.086
0.2	Overall	0.880	0.817	0.743	0.623	0.595	0.087

Outcomes (1)

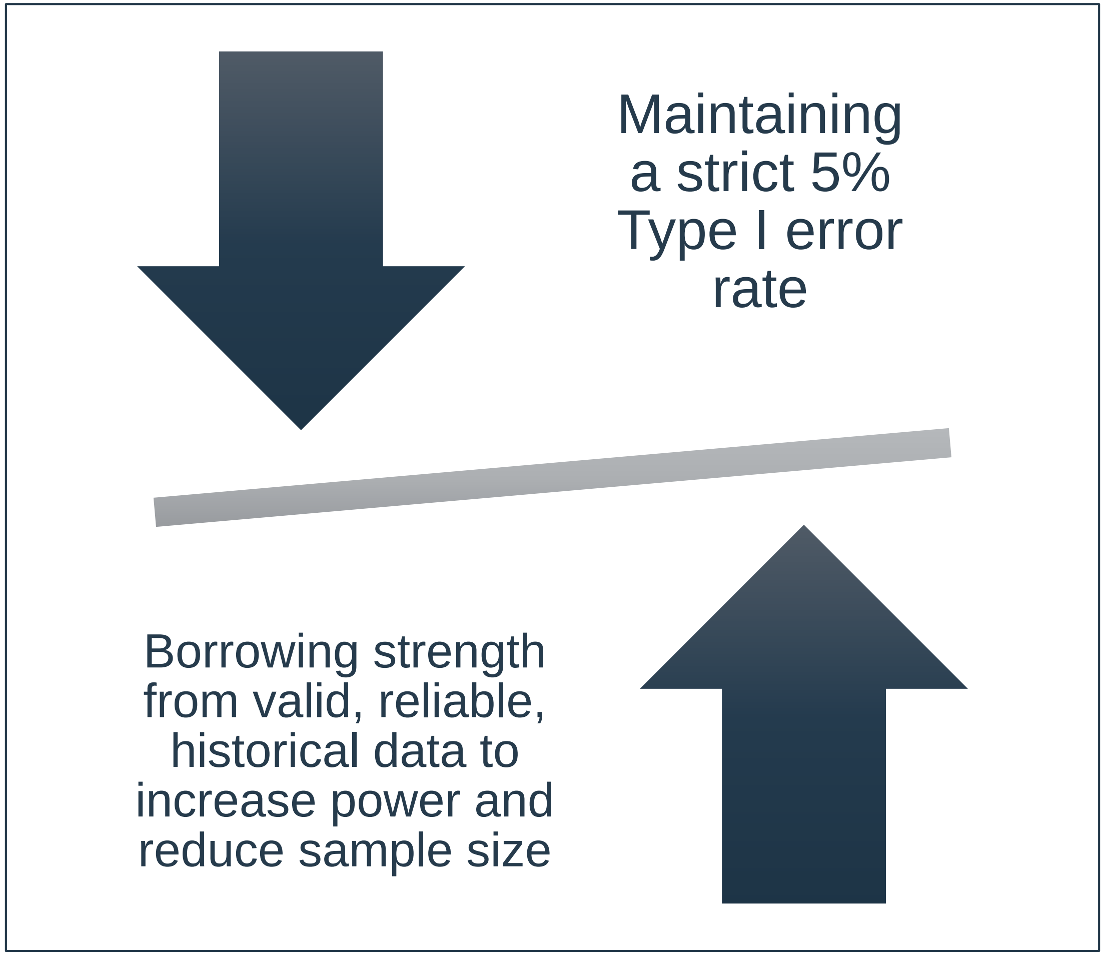
- Our design was ultimately approved by the FDA late last year, but the plan retained little Bayes advantage:
  - Bayesian analysis was relegated to secondary status– essentially forced to “match” Stouffer’s frequentist analysis, with a 5.5% conditional two-sided Type I error rate
  - The regulators’ “COVID mulligan” was much less valuable than we’d hoped or anticipated
- The approved plan provides an estimated ~80% power to detect a hazard ratio of 0.70 in the new trial given the enrollment of **2000** patients and observation of **240** events.
  - This plan incorporates weighting the previous trial’s data at 2.5% of the new data.
  - Study 1 had 406 events, so this roughly corresponds to borrowing just 406(.025) = **10** events, or just 10/250 = **4%** of the total – very conservative!

Outcomes (2)

- Our experience suggests that effective use of Bayesian adaptive design in regulatory science awaits consideration by regulators of the benefit-risk tradeoffs between Type I error inflation arising from historical data borrowing and the consequent increase in power.
  - Type I error control and p-values cannot be the only metric regulators will consider
    - Frank Harrell, FDA/Vanderbilt U: https://www.fharrell.com/post/pvalprobs/
  - 'Good data' must include more than just unseen data[2]…
  - Effective use of real world evidence (RWE) in regulatory decisions may be similarly impacted.
    - Matters are even more complicated by the need for propensity matching or other causal inference techniques

Recommendations (1)

- P-values and Type I error rates are not the only characteristics important in the design and analysis of a clinical trial, especially when incorporating external data sources.
  - The use of an informative prior increases the Type 1 error rate, by definition.
  - A 'give-and-take' or cost-benefit approach is required during regulatory decision making with external data borrowing.



Maintaining a strict 5% Type I error rate

Borrowing strength from valid, reliable, historical data to increase power and reduce sample size

Recommendations (2)

- We actively encourage continued intra-agency and cross-disciplinary statistical education focused on the evaluation and interpretation of complex innovative [trial] designs (CID).
  - FDA Statisticians could better educate program directors…
  - …and Bayesian statisticians could attend more medical meetings!
  - Bayesian design and analysis is increasingly considered in drug trials to enhance clinical trial efficiency using CIDs[3], especially in rare and pediatric disease
  - Knowledge dissemination regarding impacts of borrowing will be critical to achieve that goal.

References

1. Carlin, Bradley P., and Thomas A. Louis. Bayesian methods for data analysis. CRC press, 2009.

2. Viele, Kert, et al. "Use of historical control data for assessing treatment effects in clinical trials." Pharmaceutical statistics 13.1 (2014): 41-54. https://doi.org/10.1002/pst.1589

3. Price, Dionne, and John Scott. "The US food and drug administration’s complex innovative trial design pilot meeting program: progress to date." Clinical trials 18.6 (2021): 706-710. https://doi.org/10.1177/17407745211050580