

---

# Adaptive Group Sequential Trials with Multiple Endpoints

FDA and Industry Workshop, Washington DC  
September 15, 2008

---

Cyrus R. Mehta, Ph.D  
Cytel Inc., Cambridge, MA

email: [mehta@cytel.com](mailto:mehta@cytel.com) – web: [www.cytel.com](http://www.cytel.com) – tel: 617-661-2011

# Acknowledgements

---

- Collaborations with Ping Gao and James Ware on sample size re-estimation
- Collaborations with Ajit Tamhane and Lingyun Liu on secondary endpoints

# Motivating Example

---

- Two-arm trial of Fibromyalgia
- Primary endpoint: Pain on Numeric Rating Scale (NRS)
- Secondary endpoint: Quality of Life (QLS)
- Uncertainty about effect sizes for both endpoints
- Design for **early stopping** and **sample size increase**, while **adjusting for multiplicity** caused by two endpoints

# Statement of Problem

---

- Two stage design with sample sizes:  
 $n_1$  for stage 1;  $n_2$  for stage 2;  $n_T = n_1 + n_2$  total
- Two endpoints:  
primary  $\sim N(\mu, \sigma^2)$  and secondary  $\sim N(\xi, \tau^2)$
- Unblinded look at end of stage 1
  - Possible early stopping for efficacy or futility
  - Possible sample size increase
- Devise a testing strategy for both primary and secondary endpoints that controls the family-wise error rate (FWER)

# What is Already Known?

---

- If no secondary endpoint, Cui, Hung and Wang (CHW), (1999) weighted statistic will control type-1 error
- If no sample size increase and no early efficacy stopping, fixed-sequence tests (Weins, 2003) will control type-1 error
  - test primary at level  $\alpha$
  - test secondary at level  $\alpha$  only if primary rejects
- In all other cases special procedures are required

# CHW Method for Primary Endpoint

- **Classical statistics** if no sample size change at interim

- stage-one statistic:  $X_1 = \frac{\hat{\mu}_1 \sqrt{n_1}}{\sigma}$
- stage-two statistic:  $X_2 = \frac{\hat{\mu}_2 \sqrt{n_2}}{\sigma}$
- combined statistic:  $X_T = \frac{\hat{\mu}_T \sqrt{n_T}}{\sigma} = \sqrt{\frac{n_1}{n_T}} X_1 + \sqrt{\frac{n_2}{n_T}} X_2$

- **Classical statistics** if sample size increases from  $n_T$  to  $n_T^* = n_1 + n_2^*$

- stage-one statistic:  $X_1 = \frac{\hat{\mu}_1 \sqrt{n_1}}{\sigma}$
- stage-two statistic:  $X_2^* = \frac{\hat{\mu}_2^* \sqrt{n_2^*}}{\sigma}$
- combined statistic:  $X_T^* = \frac{\hat{\mu}_T^* \sqrt{n_T^*}}{\sigma} = \sqrt{\frac{n_1}{n_T^*}} X_1 + \sqrt{\frac{n_2^*}{n_T^*}} X_2^*$

- **CHW statistic** if sample size increases from  $n_T$  to  $n_T^*$

- $X_{\text{CHW}}^* = \sqrt{\frac{n_1}{n_T}} X_1 + \sqrt{\frac{n_2}{n_T}} X_2^* \neq \frac{\hat{\mu}_T^* \sqrt{n_T^*}}{\sigma}$
- Use  $X_{\text{CHW}}^*$  instead of  $X_T^*$  protects type-1 error

# Drawbacks of $X_{CHW}^*$ Statistic

---

- Not a function of  $\hat{\mu}^*$ , hence violates the sufficiency principle; some efficiency loss (Tsiatis and Mehta, 2003)
- Uses pre-specified weights that down-weight the second-stage estimate of  $\mu$
- P-values, unbiased point-estimates and confidence intervals require more complex statistical procedures
- No obvious way to control FWER with multiple endpoints
- $X_T^*$  has none of these drawbacks. But can it control alpha?

# Replacing CHW Statistic with Classical Statistic

---

Replace  $X_{\text{CHW}}^*$  with  $X_T^*$ , and compensate by adjusting the critical region (Gao, Ware and Mehta, 2008)

$$\text{CHW Test: } P_0(X_1 \geq c_1) + P_0(X_1 < c_1, X_{\text{CHW}}^* \geq c_T) = \alpha$$

$$\text{ACR Test: } P_0(X_1 \geq c_1) + P_0(X_1 < c_1, X_T^* \geq c_T(X_1)) = \alpha$$

where

$$c_T(X_1) = (n_T^*)^{-0.5} \left[ \sqrt{(n_2^*/n_1)} (C_\alpha \sqrt{n_T} - X_1 \sqrt{n_1}) + X_1 \sqrt{n_1} \right]$$

# Using the Classical Test without Any Adjustment

---

- Identify the range of  $X_1$  values, say  $\mathcal{R}$ , for which  $c_T(X_1) \leq c_T$
- For all  $X_1 \in \mathcal{R}$  we can increase sample size without any adjustment since

$$P_0(X_1 \geq c_1) + P_0(X_1 < c_1, X_T^* \geq c_T) \leq$$
$$P_0(X_1 \geq c_1) + P_0(X_1 < c_1, X_T^* \geq c_T(X_1)) = \alpha$$

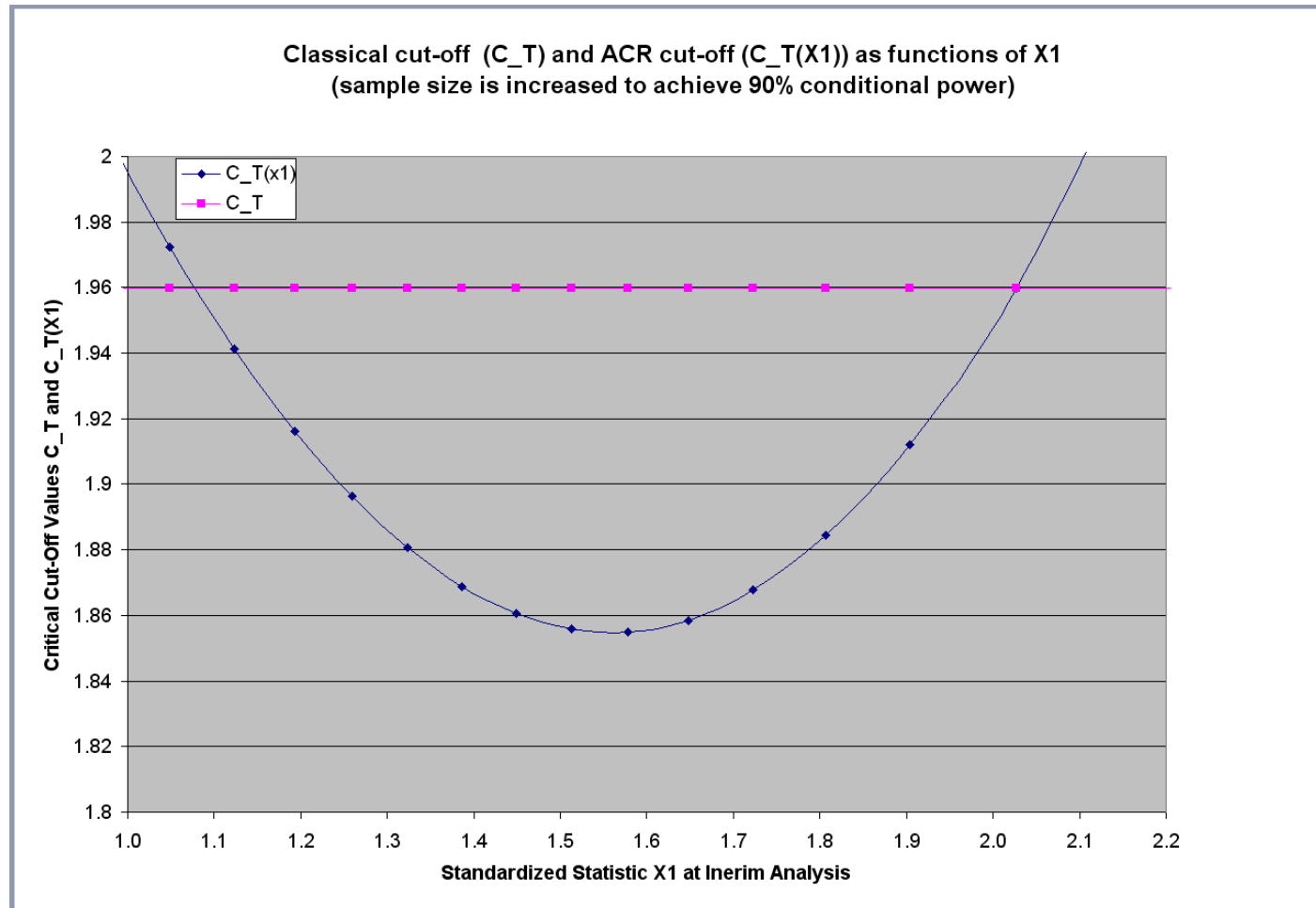
- Extends result by Chen, DeMets and Lan (2004) but  $\mathcal{R}$  covers a much broader range of  $X_1$  values

# Fibromyalgia Example

---

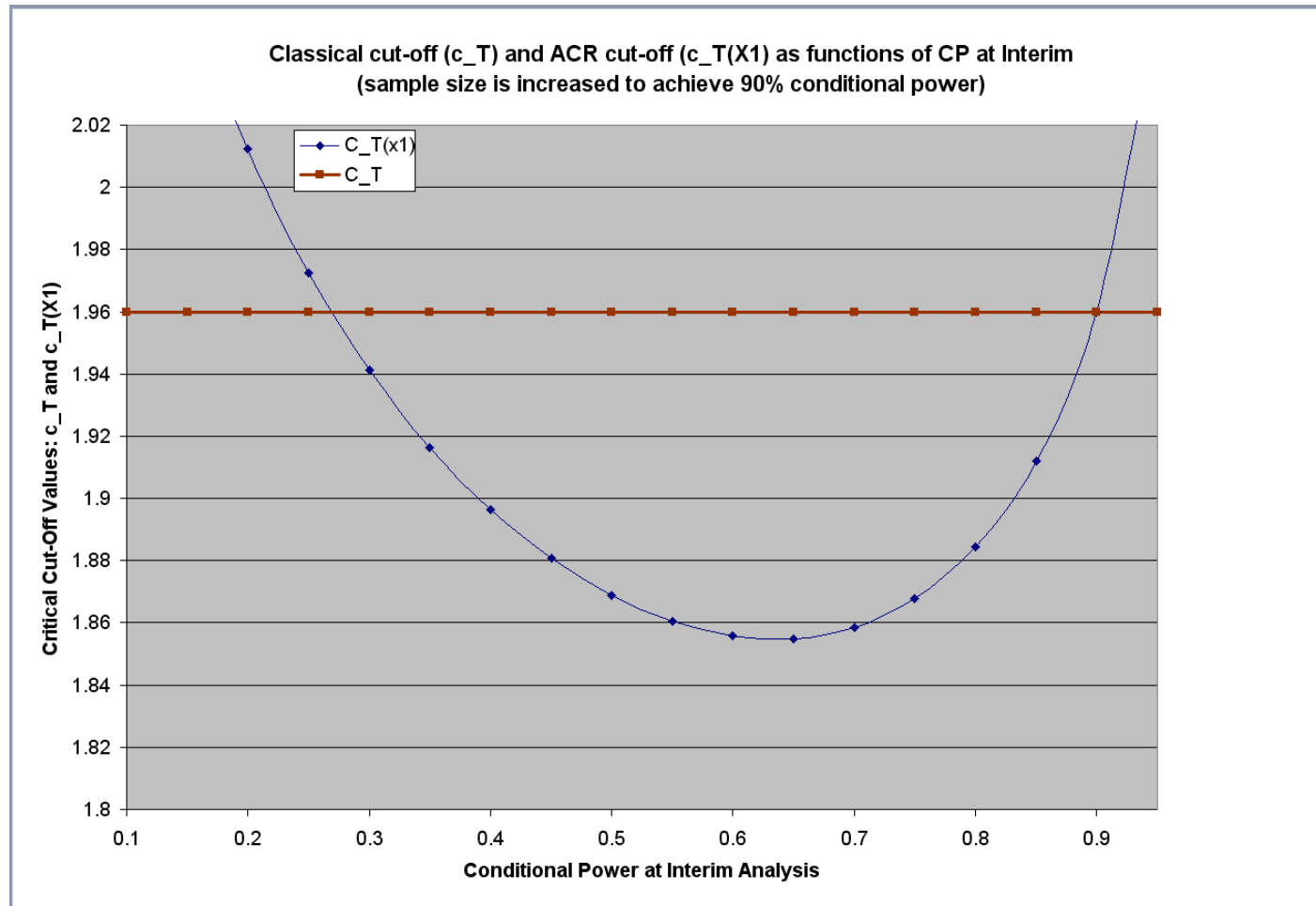
- Primary endpoint: decrease in NRS pain score at week-12
- Planned enrollment 200 subjects (100/arm) to achieve 90% power
- Increase sample size at interim to achieve 90% conditional power, **up to maximum 400 subjects**

# Region $\mathcal{R}$ in Terms of Interim $X_1$



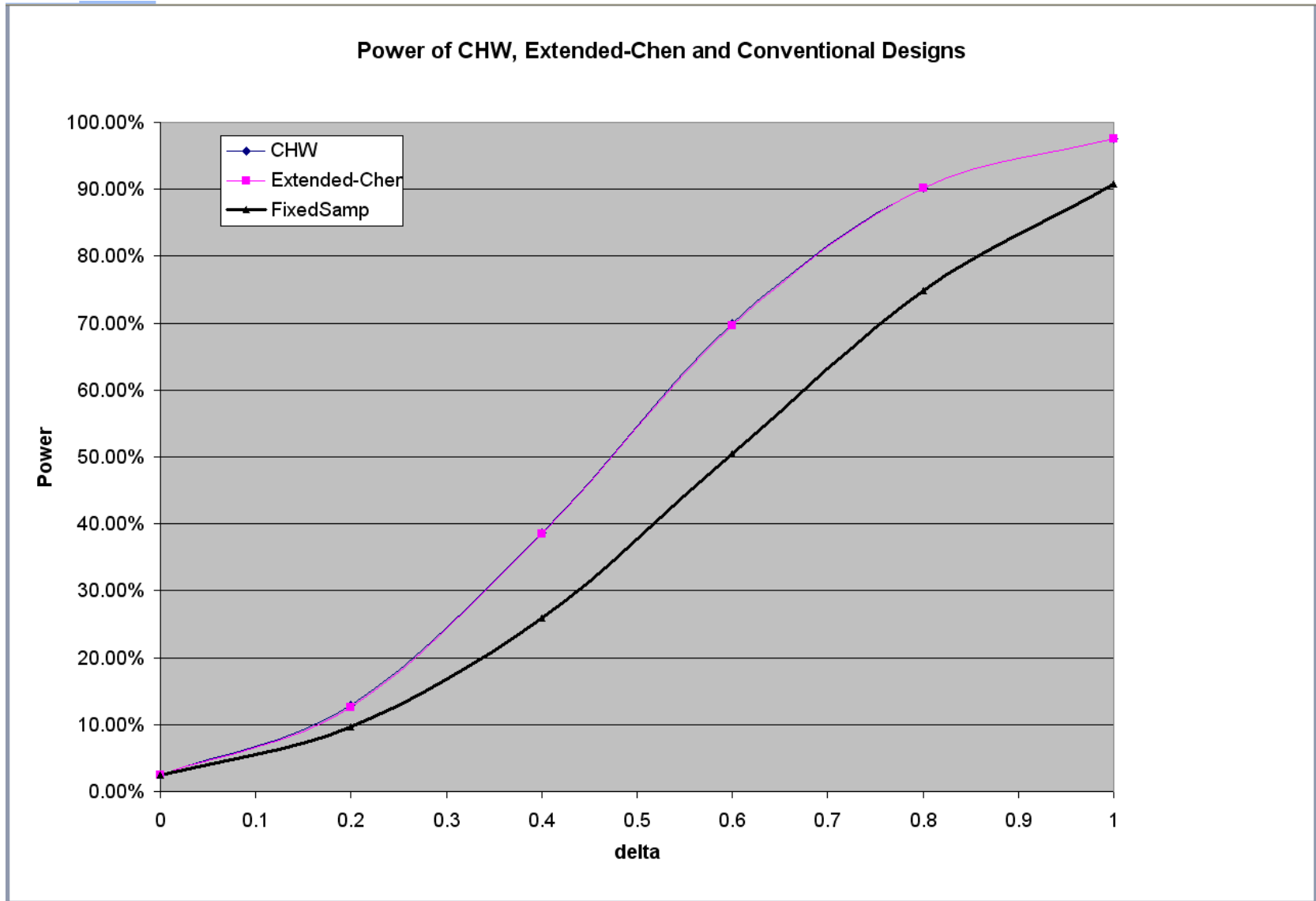
$\mathcal{R} = \{X_1: 1.1 \leq X_1 \leq 2\}$ . If  $X_1 \in \mathcal{R}$  use classical test regardless of sample size increase (extends Chen et. al. range)

# Region $\mathcal{R}$ in Terms of Interim CP



$\mathcal{R} = \{CP: 0.28 \leq CP \leq 0.9\}$ . If  $CP \in \mathcal{R}$  use classical test regardless of sample size increase. (extends Chen et. al. range)

# Negligible Power Loss Relative to CHW Test



# Final Comments on Sample Size Re-estimation

---

- CHW test is usually recommended if there is a data dependent sample size increase
- Use of the CHW statistic can be inconvenient for a variety of reasons
- One can instead use the traditional Wald statistic, with no inflation of alpha and negligible power loss, for a wide range of interim results

# Handling Secondary Endpoints

- Primary endpoint  $\sim N(\mu, \sigma)$

– stage-one statistic:  $X_1 = \frac{\hat{\mu}_1 \sqrt{n_1}}{\sigma} \sim N\left(\frac{\hat{\mu}_1 \sqrt{n_1}}{\sigma}, 1\right)$

– final statistic:  $X_T = \frac{\hat{\mu}_T \sqrt{n_T}}{\sigma} \sim N\left(\frac{\hat{\mu}_T \sqrt{n_T}}{\sigma}, 1\right)$

- Secondary endpoint  $\sim N(\xi, \tau)$

– stage-one statistic:  $Y_1 = \frac{\hat{\xi}_1 \sqrt{n_1}}{\tau} \sim N\left(\frac{\hat{\xi}_1 \sqrt{n_1}}{\tau}, 1\right)$

– final statistic:  $Y_T = \frac{\hat{\xi}_T \sqrt{n_T}}{\tau} \sim N\left(\frac{\hat{\xi}_T \sqrt{n_T}}{\tau}, 1\right)$

- $\text{corr}(X_1, Y_1) = \text{corr}(X_T, Y_T) = \rho$
- This notation will suffice even if sample size is increased at interim, by applying preceding results

# Familywise Error Rate (FWER)

---

- Let  $H_\mu: \mu = 0$  and  $H_\xi: \xi = 0$  denote the null hypotheses for the primary and secondary endpoints
- $H_\mu$  tested first.  $H_\xi$  tested **only if**  $H_\mu$  rejects
- Let  $(c_1, c_T)$  and  $(d_1, d_T)$  be the critical values at the two looks for testing  $H_\mu$  and  $H_\xi$  against one sided alternatives

$$\text{FWER} = P(X_1 \geq c_1, Y_1 \geq d_1) + P(X_1 < c_1, X_T \geq c_T, Y_T \geq d_T)$$

- Find  $(c_1, c_T)$  and  $(d_1, d_T)$  such that  $\text{FWER} \leq \alpha$
- It suffices to protect FWER only for the case  $\mu \geq 0, \xi = 0$

# Level- $\alpha$ Fixed Sequence Tests

---

- Let  $(c_1, c_T)$  be critical constants for primary hypotheses:

$$P_0(X_1 \geq c_1) + P_0(X_1 < c_1, X_T \geq c_T) = \alpha$$

- Suppose there is no interim analysis ( $c_1 = \infty, c_T = z_\alpha$ ). If  $(d_1 = \infty, d_T = z_\alpha)$ , then  $\text{FWER} \leq \alpha$ ; Weins, (2003)
- It is tempting to try the same strategy when there is interim analysis; i.e., set  $d_1 = d_2 = z_\alpha$
- Hung, Wang and O'Neil (2007) have shown through simulation this strategy fails;  $\text{FWER} > \alpha$

# Expressions for FWER

- Under the configuration  $\mu > 0, \xi = 0$

$$\begin{aligned} \text{FWER} = & \int_{-\infty}^{\infty} \Phi \left[ \frac{\delta_1 - c_1 + \sqrt{\rho}u}{\sqrt{1-\rho}} \right] \Phi \left[ \frac{-d_1 + \sqrt{\rho}u}{\sqrt{1-\rho}} \right] \phi(u) du \\ & + \int_{c_T - \delta_T}^{\infty} \Phi \left[ \frac{c_1 - \delta_1 - \gamma u}{\sqrt{1-\gamma^2}} \right] \Phi \left[ \frac{-d_T + \rho u}{\sqrt{1-\gamma^2}} \right] \phi(u) du \end{aligned}$$

where  $\delta_1 = \frac{\mu\sqrt{n_1}}{\sigma}$ ,  $\delta_T = \frac{\mu\sqrt{n_T}}{\sigma}$  and  $\gamma = \sqrt{\frac{n_1}{n_T}}$

- For the special case of  $\rho = 1$

$$\begin{aligned} \text{FWER} = & \Phi\{-\max(c_1 - \delta_1; d_1)\} \\ & + \int_{-\infty}^{c_1 - \delta_1} \Phi\left\{ \frac{y\sqrt{n_1} - \max(c_T - \delta_T, d_T)\sqrt{n_T}}{\sqrt{n_T - n_1}} \right\} \phi(y) dy \end{aligned}$$

# Fibromyalgia Example Re-visited

---

- Primary endpoint is NRS. Secondary endpoint is QLS
- Suppose treatment is efficacious for primary endpoint (NRS) but no different from placebo for secondary endpoint (QLS); i.e.,  $\mu \geq 0, \xi = 0$
- Evaluate FWER for various boundaries  $(c_1, c_T)$  for the primary endpoint and  $(d_1, d_T)$  for the secondary boundary

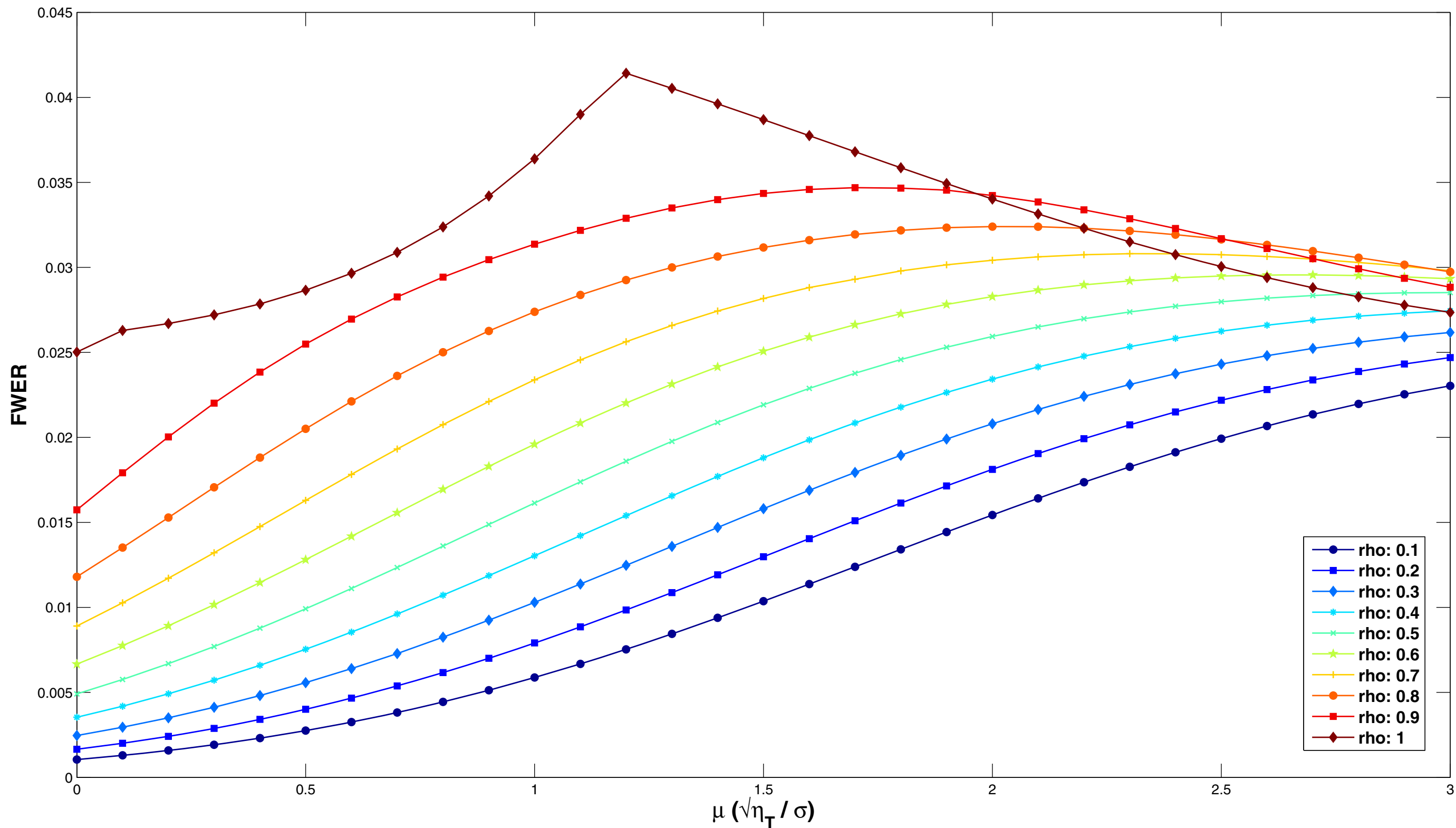
# Result 1

---

- Suppose  $(c_1, c_T)$  are the boundaries for a level- $\alpha$  test of the primary endpoint. Then  $d_1 = d_T = z_\alpha$  for the secondary endpoint will inflate the FWER
- This result was shown by Hung, Wang and O'Neill (2007) using simulation
- **Example:**  
 $(c_1, c_T) = (2.797, 1.977)$ ; O'Brien-Fleming level-0.025 boundaries  
 $(d_1, d_T) = (1.96, 1.96)$ ; spend  $\alpha = 0.025$  for secondary endpoint as soon as primary crosses boundary

# O'Brien-Fleming primary, Level 0.025 secondary at each look

$d_1 = 1.96; d_T = 1.96; c_1 = 2.797; c_T = 1.977$



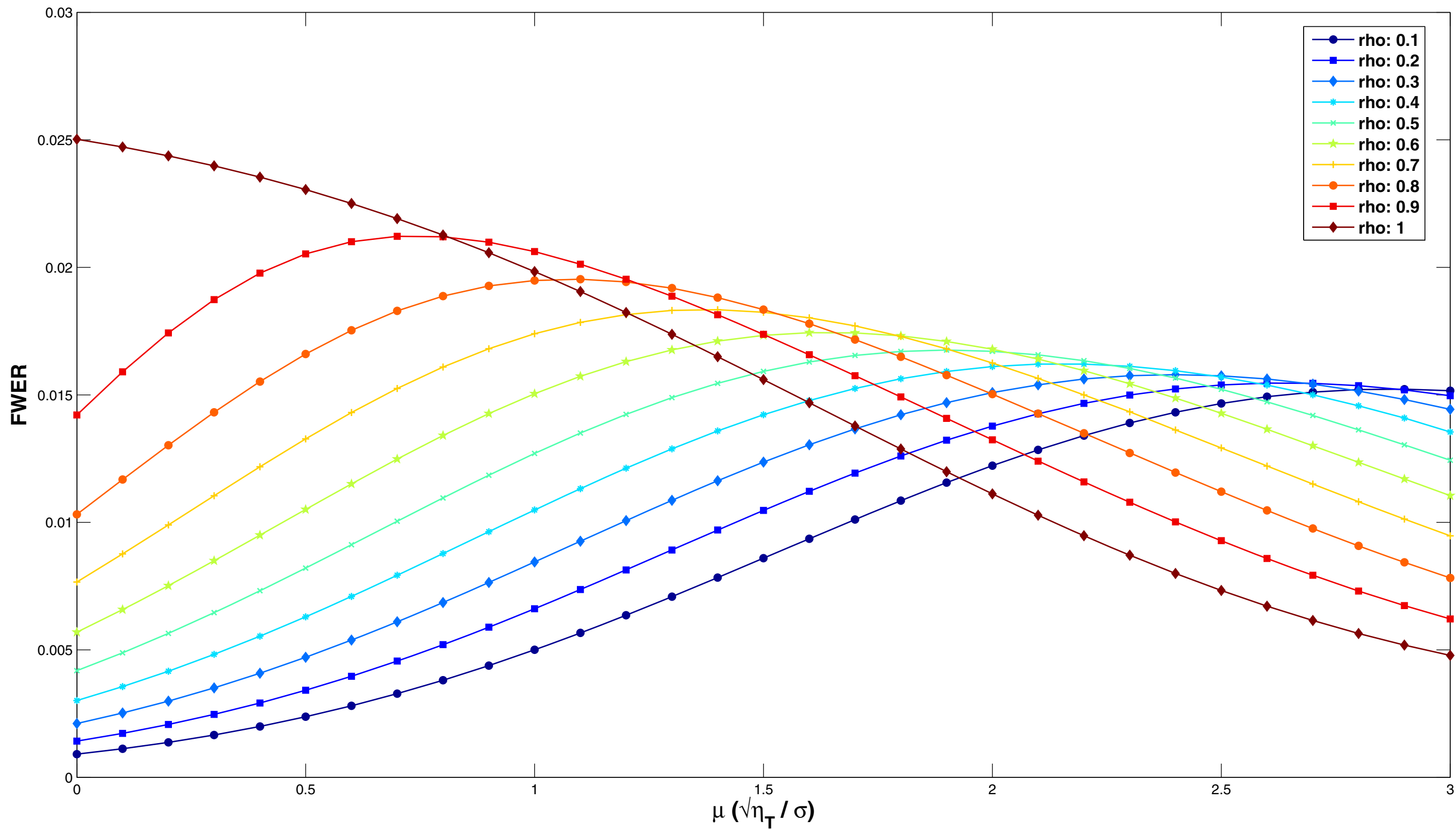
# Result 2

---

- If  $c_1 = d_1$  and  $c_T = d_T$  FWER is preserved
- **Example (a):**
  - $(c_1, c_T) = (2.797, 1.977)$ ; O'Brien-Fleming primary
  - $(d_1, d_T) = (2.797, 1.977)$ ; O'Brien-Fleming secondary
- **Example (b):**
  - $(c_1, c_T) = (2.178, 2.178)$ ; Pocock primary
  - $(d_1, d_T) = (2.178, 2.178)$ ; Pocock secondary

# O'Brien–Fleming primary, O'Brien–Fleming secondary

$d_1 = 2.797$ ;  $d_T = 1.977$ ;  $c_1 = 2.797$ ;  $c_T = 1.977$





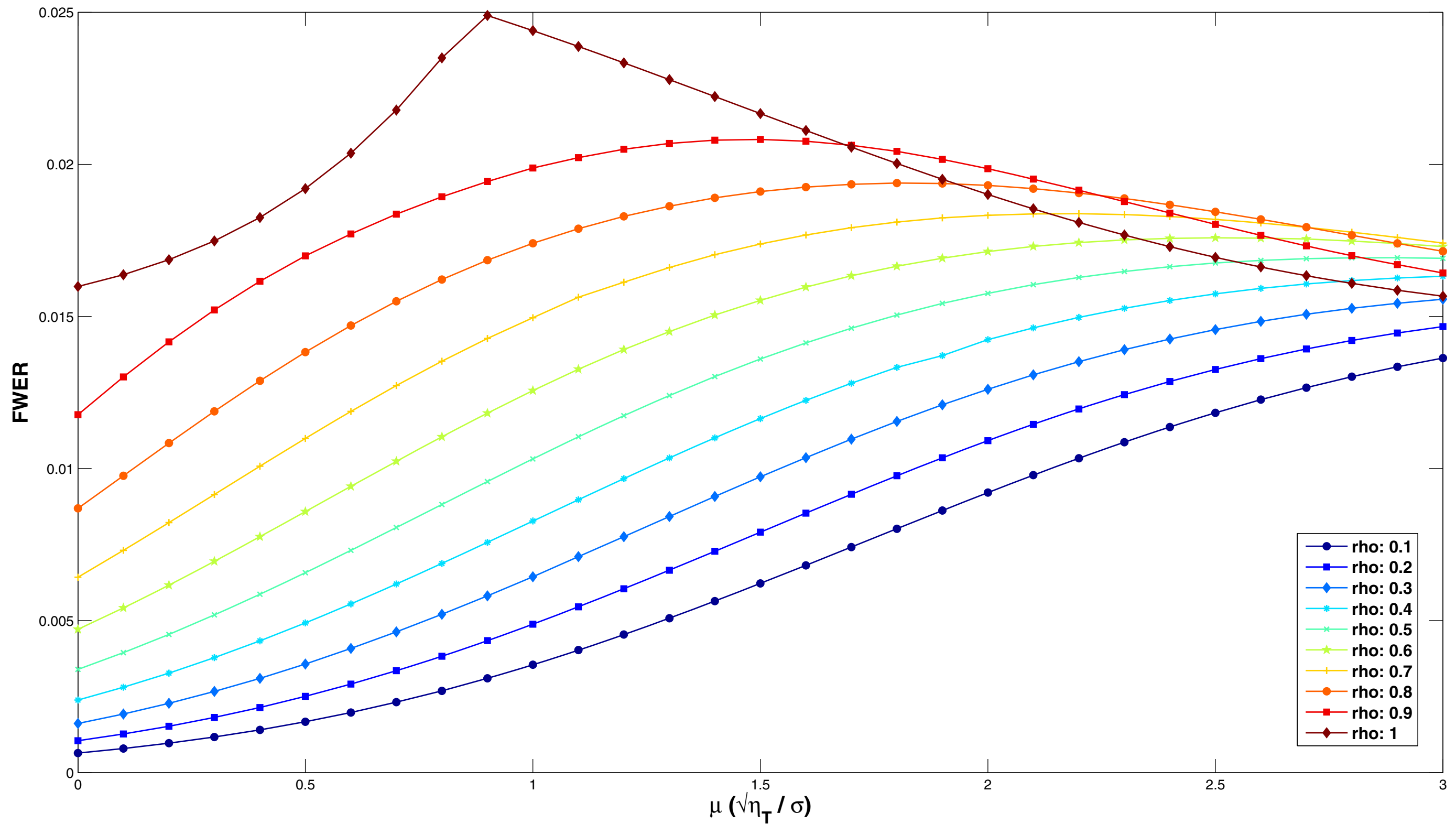
# Result 3

---

- If  $(c_1, c_T)$  and  $(d_1, d_T)$  are any level- $\alpha$  boundaries, and primary spends  $\alpha$  **slower** than secondary,  $\text{FWER} = \alpha$
- **Example:**  
 $(c_1, c_T) = (2.797, 1.977)$ ; O'Brien-Fleming primary  
 $(d_1, d_T) = (2.178, 2.178)$ ; Pocock secondary

# O'Brien–Fleming primary, Pocock secondary

$d_1 = 2.178; d_T = 2.178; c_1 = 2.797; c_T = 1.977$



# Result 4

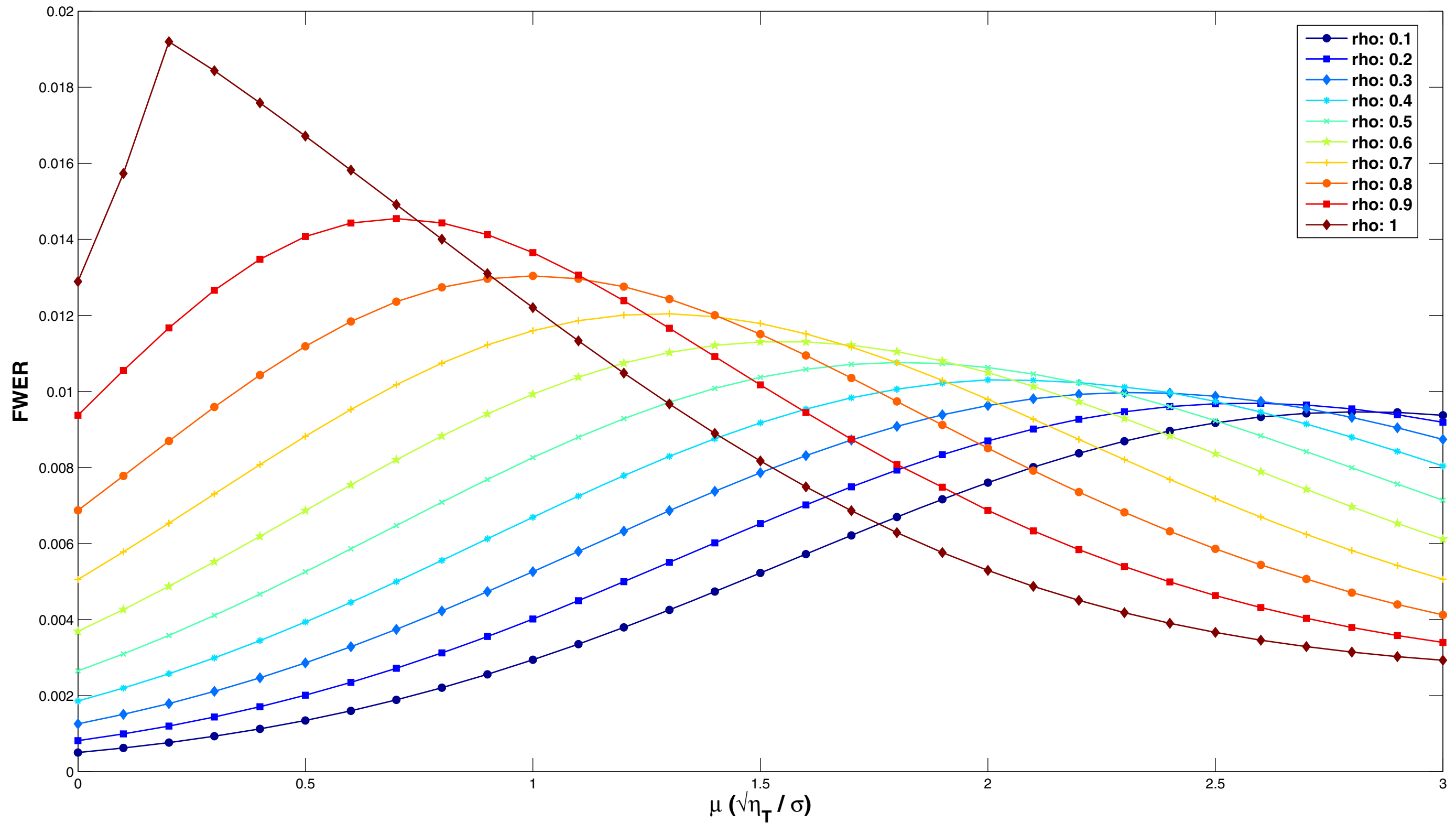
---

- If  $(c_1, c_T)$  and  $(d_1, d_T)$  are any level- $\alpha$  boundaries, and primary spends  $\alpha$  **faster** than secondary,  $\text{FWER} < \alpha$
- **Example:**
  - $(c_1, c_T) = (2.178, 2.178)$ ; Pocock primary
  - $(d_1, d_T) = (2.797, 1.977)$ ; O'Brien-Fleming secondary

This gives us the opportunity to spend **more** than  $\alpha$  for the secondary and still preserve the FWER

# Pocock primary; O'Brien–Fleming secondary

$d_1 = 2.797$ ;  $d_T = 1.977$ ;  $c_1 = 2.178$ ;  $c_T = 2.178$



# Final Conclusions

---

- Worst case arises at  $\rho = 1$
- As long as primary and secondary boundaries are both at level- $\alpha$ , the FWER is guaranteed not to exceed  $\alpha$
- If primary spends the  $\alpha$  faster than secondary,  $\text{FWER} < \alpha$  and we can choose boundaries for the secondary hypothesis with error rates that exceed  $\alpha$
- These results hold even with sample size increase at interim, provided the sample size is only increased when “extended-Chen” condition is met