Randomized Clinical Trials – Replacing Traditional Analyses with Better Alternatives

Devan V. Mehrotra
Biostatistics and Research Decision Sciences

2018 ASA Traveling Short Course

ASA Kansas/Western Missouri Chapter

Oct 26, 2018

Course Agenda

9:05 – 9:10	Course Introduction
9:10 – 9:35	Crossover trials with baselines
9:35 – 10:30	Trials with unequally powered primary endpoints
10:30 – 10:45	Break
10:45 – 11:15	Small trials with time-to-event endpoints
11:15 – 12:15	Pharmacogenomics in phase 2-3 trials
12:15 – 1:15	Lunch
1:15 – 2:30	Stratified Trials: binary, time-to-event endpoints
2:30 – 2:45	Break
2:45 – 3:25	Estimand-aligned primary & sensitivity analyses
3:25 – 3:30	Wrap-Up

Course Introduction

- Increasing cost pressures and unmet healthcare needs necessitate making clinical drug development cheaper and faster while maintaining strong scientific rigor

- We describe several examples of notably more powerful clinical trial analyses that enable improved operational efficiency across all phases of clinical development relative to traditional approaches

- Big-picture "5P" focus:

  Patients, Prescribers, Payers, Pricing, Product Label

Topic #1

Crossover Trials with Baselines

2x2 Crossover with Baselines

Treatment Sequence	Period 1 Baseline $E(X_1)$	Period 1 Treatment $E(Y_1)$	WASHOUT	Period 2 Baseline $E(X_2)$	Period 2 Treatment $E(Y_2)$
A→B	λ	$\lambda + \mu_A$		λ^*	$\lambda^* + \mu_B$
B→A	λ	$\lambda + \mu_B$		λ^*	$\lambda^* + \mu_A$

Goal: estimate $\delta = \mu_A - \mu_B$, test $H_{null} : \delta = 0$

Caution: analysis is impacted by (i) how X_1, X_2 are used in the analysis, and (ii) "structure" of $\sum$ =Var(X_1, Y_1, X_2, Y_2)'

$$\Sigma_{CS} = v\begin{pmatrix} 1 & \rho & \rho & \rho \\ & 1 & \rho & \rho \\ & & 1 & \rho \\ & & & 1 \end{pmatrix} \quad \Sigma_{AR(1)} = v\begin{pmatrix} 1 & \rho & \rho^2 & \rho^3 \\ & 1 & \rho & \rho^2 \\ & & 1 & \rho \\ & & & 1 \end{pmatrix} \quad \Sigma_{EP} = v\begin{pmatrix} 1 & \rho_{12} & \rho_{13} & \rho_{14} \\ & 1 & \rho_{14} & \rho_{13} \\ & & 1 & \rho_{12} \\ & & & 1 \end{pmatrix}$$

{Commonly seen $\sum$ structures in practice}

Three Competing Methods

M1: contrast is C1 = (Y1-X1)-(Y2-X2) = (Y1-Y2)-(X1-X2) (traditional)

M2: contrast is C2 = (Y1|X1)-(Y2|X2) = (Y1-Y2)-(uX1-vX2)

M3: contrast is C3 = (Y1-Y2)|(X1-X2) = (Y1-Y2)-w(X1-X2)

- u,v,w are functions of elements of $\sum$
- E(Ci|AB) – E(Ci|BA) = 2δ for i=1,2,3 (all unbiased)
- M3 has smallest variance (Mehrotra, 2014)

Contrast variances under common $\sum$ structures

	CS	AR(1)	EP
M1	$4v(1-\rho)$	$2v(1-\rho^2)\times(2-\rho)$	$4v[1-\rho_{13}-\rho_{12}+\rho_{14}]$
M2	$2v(1-\rho)\times(1+\rho^2)$	$2v(1-\rho^2)$	$2v[1-\rho_{13}-\rho_{12}(\rho_{12}+\rho_{12}\rho_{13}-2\rho_{14})]$
M3	$2v(1-\rho)$	$2v(1-\rho^2)\times(1-\frac{\rho^2}{4})$	$2v\left[1-\rho_{13}-(\rho_{12}-\rho_{14})^2(1-\rho_{13})^{-1}\right]$

SAS Codes for the Three Competing Methods

M1: change from baseline analysis (traditional)

```
PROC MIXED DATA=dataset;
  CLASSES seq prd trt subjid;
  MODEL cfb = seq prd trt;
  REPEATED prd/SUBJECT=subjid(seq) TYPE=UN;
  ESTIMATE 'trteff' trt 1 -1/CL;
RUN;
```

M2: period-specific baseline (x) is covariate

```
MODEL y = x prd x*prd trt/DDFM=KR;
```

M3: difference between baselines (xdiff) is covariate

```
MODEL y = seq prd trt xdiff xdiff*prd; [xdiff = x1-x2]
```

Illustrative Example

Source: Mehrotra (2014)

Example 1: Mean Arterial Pressure Data, Summary Statistics, and Analysis Details

	Sequence AB					Sequence BA			
	Period 1 Treatment A		Period 2 Treatment B			Period 1 Treatment B		Period 2 Treatment A	
Subject #	X_1	Y_1	X_2	Y_2	Subject #	X_1	Y_1	X_2	Y_2
1	92.4	92.0	100.3	93.7	1*	96.5	98.0	97.9	95.2
2	101.4	100.8	93.1	96.0	2*	94.9	98.3	97.7	92.1
3	95.4	86.6	91.6	83.0	3*	87.8	91.1	80.0	87.0
4	101.2	100.3	101.9	103.0	4*	93.9	89.7	91.3	87.7
5	101.2	99.9	103.9	97.8	5*	77.7	77.8	78.9	82.5
6	96.6	99.6	98.6	100.8	6*	99.2	100.0	101.0	98.0
7	103.9	98.4	98.6	100.8	7*	98.5	97.1	96.2	94.1
8	81.2	83.5	86.7	83.0	8*	84.6	83.8	82.1	80.0
9	81.0	79.3	78.1	84.0	9*	91.4	94.4	91.6	85.6
10	98.9	92.2	88.7	88.0	10*	95.3	91.7	99.4	98.6
11	79.3	76.6	78.1	75.2	11*	94.4	90.4	91.9	88.6
12	90.3	83.2	85.3	86.6	12*	89.5	89.6	75.4	82.3
13	90.3	85.4	96.8	101.1	13*	89.4	88.9	93.5	83.9
Mean	*93.3*	*90.6*	*92.4*	*91.8*	*Mean*	*91.8*	*91.6*	*90.5*	*88.9*
S.D.	*8.5*	*8.7*	*8.6*	*9.0*	*S.D.*	*6.0*	*6.2*	*8.6*	*6.2*

X_1 = baseline response in period 1, Y_1 = treatment response in period 1, etc.

Illustrative Example (continued)

AICC
(smaller is better)

$$\hat{\Sigma}_{UN} = \begin{pmatrix} 53.94 & 50.37 & 49.33 & 43.28 \\ [.91] & 56.89 & 50.81 & 46.24 \\ [.78] & [.78] & 73.86 & 55.53 \\ [.77] & [.80] & [.84] & 59.27 \end{pmatrix}$$

CS	602.5
AR(1)	605.0
EP	600.2
UN	609.0

(REML estimates; correlations in brackets)

Analysis Method		$\hat{\delta}$	SE($\hat{\delta}$)	p-value
M1	(Y1-X1)-(Y2-X2)	-1.75	1.06	.111
M2	(Y1|X1)-(Y2|X2)	-2.05	0.95	.043
M3	(Y1-Y2)|(X1-X2)	-1.86	0.87	.043

Simulation Results: Power (%)

N=16/seq, mean pairwise correlation = 0.6

Details: Mehrotra (2014)

Analysis Method			True $\sum$ structure		
			CS	AR(1)	EP
1	(Y1-X1)-(Y2-X2)	Change from baseline	50	70	80
2	(Y1|X1)-(Y2|X2)	Period-specific covariates	65	77	86
3	(Y1-Y2)|(X1-X2)	X1-X2 is a covariate	80	84	90

- Method 1 (traditional) is statistically inefficient
- Method 3 improves power, resulting in typical sample size savings of ~ 30% versus Method 1

Conclusions: Topic #1

For 2x2 crossovers with period-specific baselines:

- The analysis is impacted by (i) how are X_1, X_2 are used, and (ii) the unknown "structure" of Var(X_1,Y_1,X_2,Y_2)
- The traditional 'change from baseline' analysis is inefficient for all common covariance structures
- A much better option is to use the difference between period-specific baselines as a covariate
- Similar idea works nicely for crossovers with > 2 periods (Jemielita et al, 2016).

References: Topic #1

- Jemielita T, Putt M, Mehrotra DV (2016). Statistics in Medicine, 35, 5625-5641
- Kenward M, Roger J (2010). Biostatistics, 11, 1-17
- Mehrotra DV (2014). Pharmaceutical Statistics, 13, 376-387
- Metcalfe C (2010). Statistics in Medicine, 29, 3211-3218

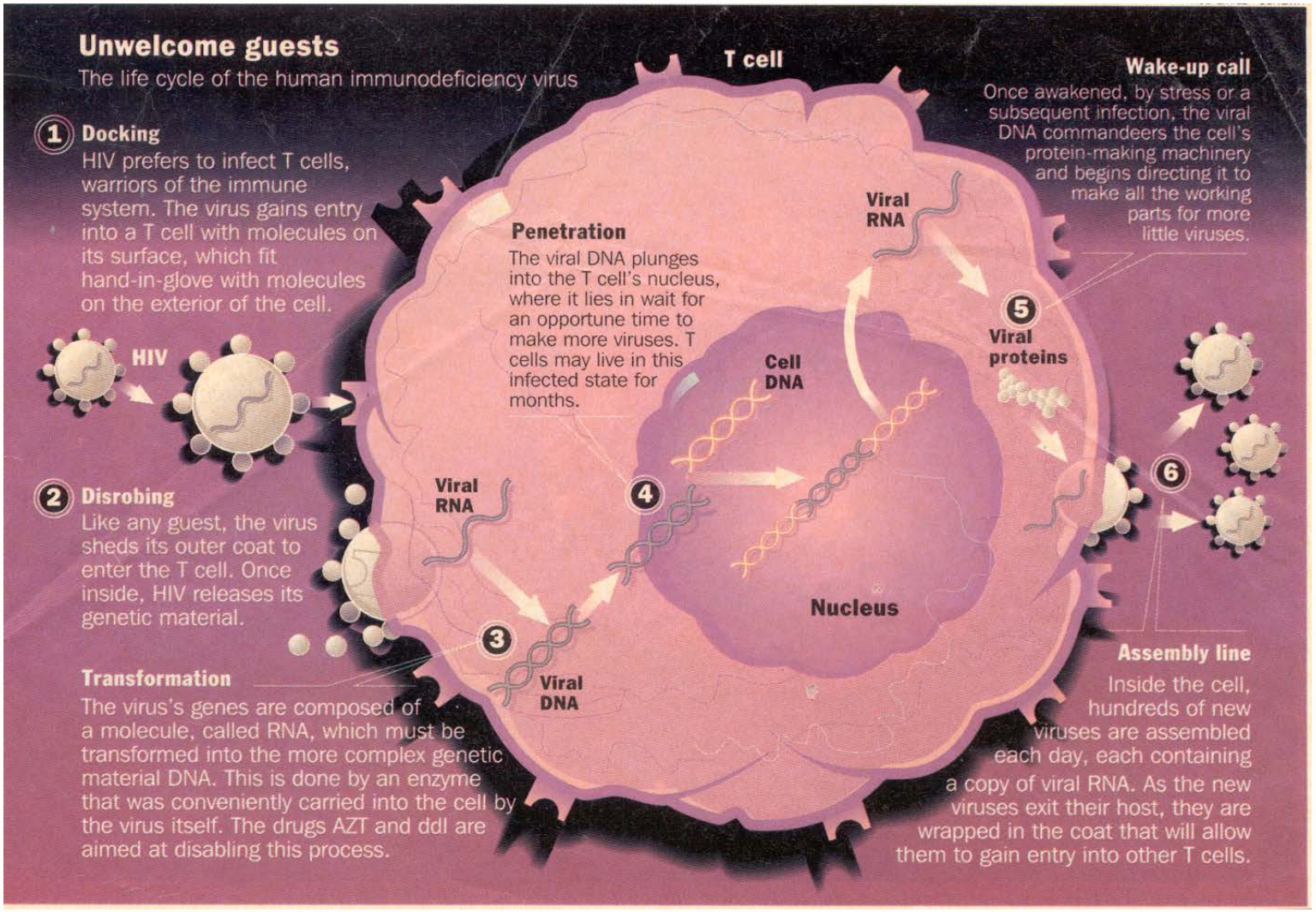
Topic #2

Trials with Unequally Powered Primary Endpoints

Case Study: HIV Vaccine Trial (2008)

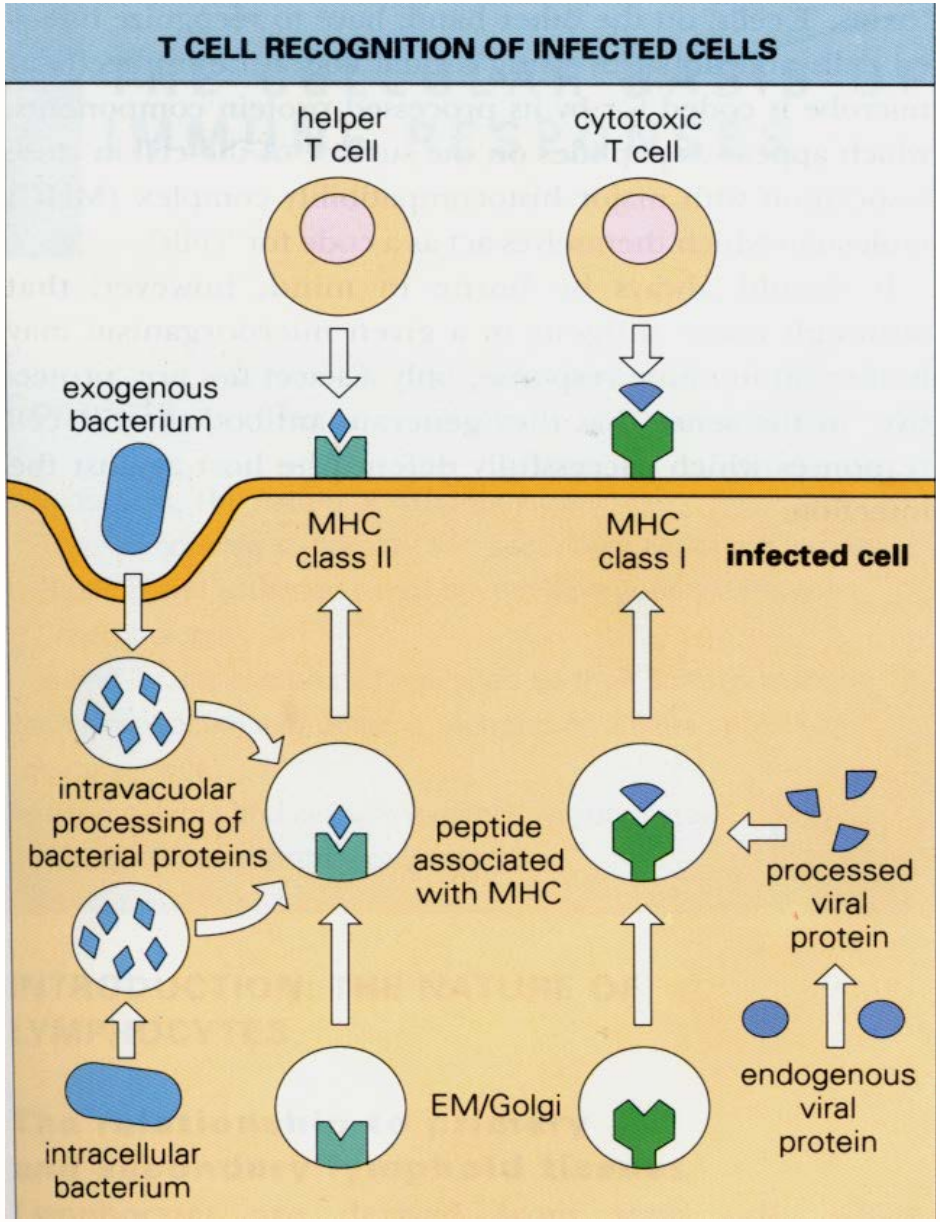
HIV Infecting a CD4 Cell

The immune system needs CD4 cells to function



Cell Mediated Immunity: Mechanism

Mims CA, Playfair JHL, Roitt IM, et al (1993). "Medical Microbiology"



CMI-based vaccine (goal): create an army of cytotoxic T cells trained to recognize and kill HIV infected cells

CMI-Based Vaccine: How to Test Efficacy?

CMI-based vaccines: large uncertainty $\Rightarrow$ prudent to conduct a focused test-of-concept (TOC) efficacy trial with a lower [but clinically relevant] hurdle relative to a traditional phase III efficacy trial

Three key TOC trial design questions:

1. Primary efficacy endpoint(s)?
2. Statistical method for efficacy evaluation?
3. How to accelerate GO/NO GO to phase III?

#1. Primary Efficacy Endpoint(s)

- For an **antibody-based** vaccine, a phase III efficacy trial would use infection (INF) as the primary endpoint

  Example: H_{null}: $VE_{INF} \leq 30\%$ vs. H_{alt}: $>30\%$

  where $VE_{INF} = 1 - \frac{\text{infection rate for VACCINE}}{\text{infection rate for PLACEBO}}$

  If true $VE_{INF} = 60\%$, for 90% power, 1-tailed α = 2.5%, **160 infections** needed to establish efficacy

Primary Efficacy Endpoint(s) (continued)

- Appropriate hypotheses for a **TOC** trial:

  H_{null}: $VE_{INF} = 0\%$ and $VE_{VL} = 0$

  H_{alt}: $VE_{INF} > 0\%$ and/or $VE_{VL} > 0$

  VE_{VL} = true difference in mean pathogen load among infected subjects (placebo – vaccine)

- **Test of concept** is successful if H_{null} can be rejected with 95% confidence
- What **statistical method** is appropriate for testing the composite efficacy hypothesis above?

#2. Statistical Method for Efficacy Evaluation

- Single analysis of a composite BOI endpoint **[Method A]**
  - Traditional method
  - Define burden-of-illness (BOI) = viral load for infected subject and zero for uninfected subject
  - Use a single statistical test to compare the BOI per randomized subject between vaccine and placebo grps
- Separate analyses of INF and VL endpoints **[Method B]**
  - Use separate statistical tests for the infection and viral load endpoints
  - Pay a statistical price for seeking two chances to establish vaccine efficacy (multiplicity adjustment to keep the false positive risk under 5%)

TOC Efficacy Trial: Data Set-Up

	Vaccine	Placebo
Number randomized	N_v	N_p
Number HIV infected	n_v	n_p
Proportion infected	$\frac{n_v}{N_v}$	$\frac{n_p}{N_p}$
Viral load set-points of infected subjects ($\log_{10}$ copies/ml)	$\begin{bmatrix} y_1^{(v)} \\ \\ y_{n_v}^{(v)} \end{bmatrix}$	$\begin{bmatrix} y_1^{(p)} \\ \\ y_{n_p}^{(p)} \end{bmatrix}$

Statistical Method for Efficacy Evaluation

- **Method A (BOI analysis of composite endpoint)**

Difference in BOI per subject: $T = \frac{\sum_{i=1}^{n_v} y_i^{(v)}}{N_v} - \frac{\sum_{i=1}^{n_p} y_i^{(p)}}{N_p}$

Let $Z_{BOI} = \frac{T - E(T \mid n_v + n_p, H_{null})}{\sqrt{Var(T \mid n_v + n_p, H_{null})}}$ (Chang et al, 1994)

Reject null if $\Pr(Z < Z_{BOI} \mid H_0) < .05$ [randomized]

- **Method B (separate analyses of INF and VL endpoints)**

p_1 =p-value for infection endpoint (Binomial) [randomized]

p_2 =p-value for VL endpoint (rank test) [non-randomized]

Reject H_{null} if $\max(p_1, p_2) < .05$ **or** $\min(p_1, p_2) < .025$

Is Method A versus Method B a fair comparison?

Method B: Adjusting for Potential Selection Bias

- Test for **viral load** component in **Method B**:
  - Is restricted to subjects that are selected based on a post-randomization outcome (HIV infection)
  - Can be confounded with the effect of variables associated with VL that are unevenly distributed among the infectees $\Rightarrow$ **selection bias** problem
- A fairer comparison is Method A versus **adjusted** Method B, where the adjustment is to account for potential **selection bias** in the VL comparison

Adjusting for Selection Bias

- Proposed approach:
  - **Adjust** the viral load test for plausible levels of selection bias such that rejection of the null hypothesis becomes harder.
  - If the **adjusted** test is significant, then we have robust evidence of a causal vaccine effect.
  - The adjustment is derived via the principal stratification framework of causal inference.

Key References

Frangakis and Rubin (2002)

Hudgens, Hoering, Self (2003)

Gilbert, Bosch, Hudgens (2003) [GBH]

Mehrotra, Li, Gilbert (2006)

Shepherd, Gilbert, Mehrotra (2007)

Method A vs. Adjusted Method B

No. of Infections Needed for TOC

α=5%, 80% power

VE_{VL}	VE_{INF}	Unadjusted Method B	Adjusted* Method B	Method A
1.0	0%	28	30	> 250
1.0	30%	27	34	74
0	60%	47	48	44

* adjusted for potential selection bias in viral load comparison

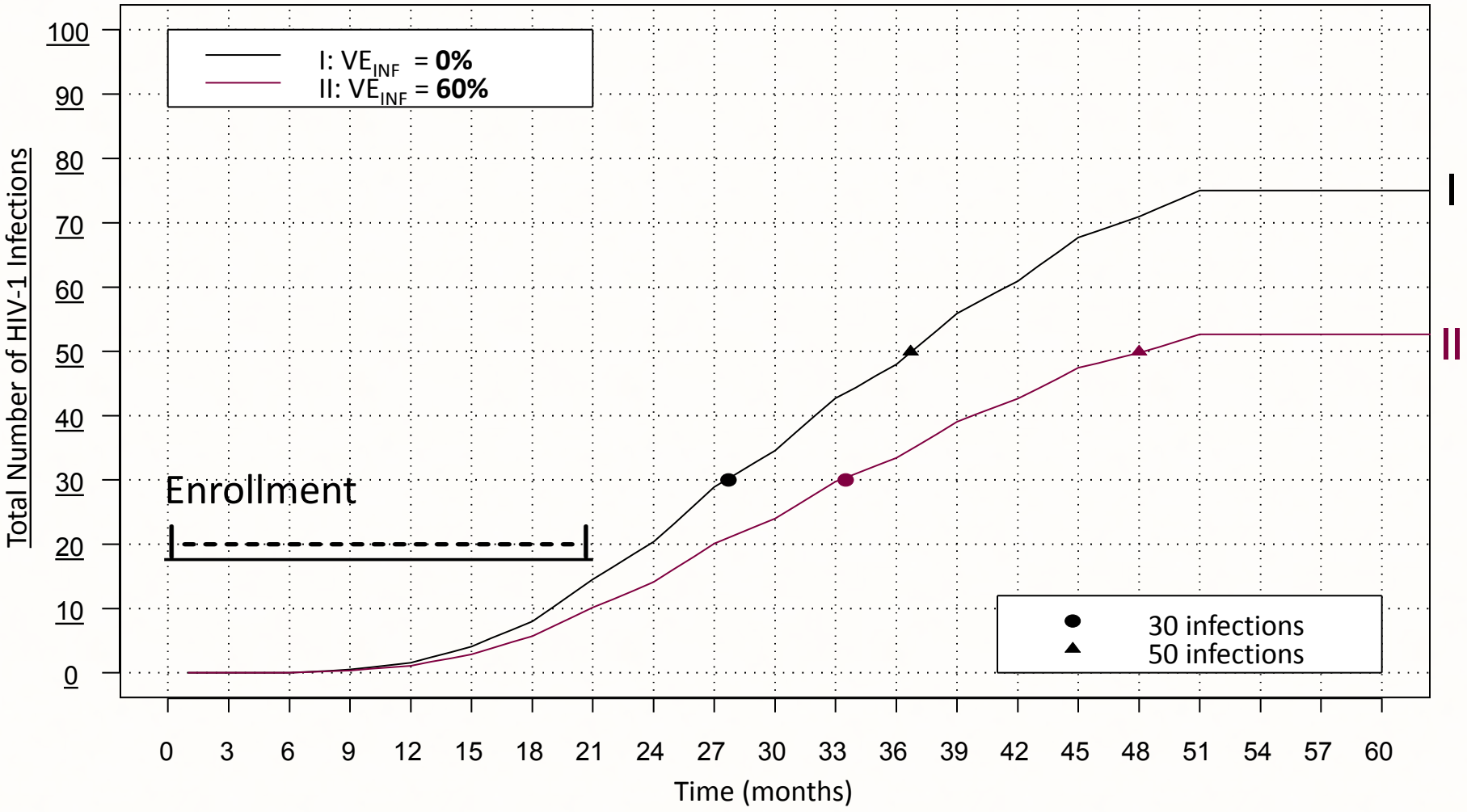
Method B (separate analysis of INF and VL endpoints) is **notably more powerful** than traditional method A (single analysis of BOI endpoint)

~ **30** infections required if VE_{INF} ~ 0% and VE_{PL} ~ 1.0 [more likely]
~ **50** infections required if VE_{INF} ~ 60% and VE_{PL} ~ 0 [less likely]

#3. Accelerating GO/NO GO for phase III

Interim analysis when viral load endpoint has 80% power

(**at 30 per-protocol HIV infections** assuming VE_{VL}=1.0 c/ml)



Above example: N = 750 per arm, ~ 21 months enrollment, placebo infection rates of ~2% per year, and dropout rates of 10%, 5%, 5% for 1st, 2nd, 3rd year, respectively.

Summary of TOC Trial Design for CMI Vaccines

- Infection and viral load are **dual primary endpoints**; success if efficacy > 0 for either endpoint
- **Separate analysis** of each endpoint (with selection bias adjustment for VL) instead of a single analysis of "burden-of-illness" that combines the two endpoints
- **Interim analysis** when viral load comparison is well powered (30 HIV infections): speeds up phase III GO/NO-GO by ~1 yr

Implementation of TOC Trial Design

- **Step** trial of MRKAd5 gag/pol/nef vs. placebo
- Initiated Dec 2004, in high risk HIV- volunteers
- Multinational: Americas, Australia, Caribbean

Step Trial: Outcome of Interim Analysis (2007)

- No evidence of vaccine efficacy
  - 30 HIV infections: 19 vaccine, 11 placebo
  - Post-infection viral loads similar for vaccine and placebo
  - **NO GO** to phase III
- The Step results recalibrated scientific discourse on the utility of CMI-based HIV vaccines, use of viral vectors to deliver vaccine antigens, etc.
- The Merck HIV vaccine failed, but the TOC trial design was hailed as a "success" for providing a NO GO in a resource-efficient manner

References: Topic #2

- Buchbinder S, Mehrotra DV et al. (2008). Efficacy assessment of a cell-mediated immunity HIV-1 vaccine (the Step Study): a double-blind, randomised, placebo-controlled, test-of-concept trial. Lancet, 372, 1881-1893.
- Chang MN, Guess HA, Heyse JF (1994). Reduction in the burden of illness: a new efficacy measure for prevention trials. Statistics in Medicine, 13, 1807-1814.
- Frangakis CE, Rubin DB (2002). Principal stratification in causal inference. Biometrics, 58, 21-29.
- Gilbert PB, Bosch RJ, Hudgens MG (2003). Sensitivity analysis for the assessment of causal vaccine effects on viral load in HIV vaccine clinical trials. Biometrics, 59, 531-541.
- Hudgens MG, Hoering A, Self SG (2003). On the analysis of viral load endpoints in HIV vaccine trials. Statistics in Medicine, 22, 2281-2298.
- Mehrotra DV, Li X, Gilbert PB (2006). A comparison of eight methods in the dual-endpoint evaluation of vaccine efficacy in a proof-of-concept HIV vaccine trial. Biometrics, 62, 893-900.
- Shepherd BE, Gilbert PB, Mehrotra DV (2007). Eliciting a counterfactual sensitivity parameter. The American Statistician, 61, 1-8.

Topic #3

Small Trials with Time-to-Event Endpoints

Two Group Time-to-Event Trial

- Subjects randomized to treatment A or B
- Time-to-event (e.g., time to death) is measured
- Proportional hazards assumption: $h_j(t) = h_0(t)e^{\beta Z_j}$

$$Z_j = \begin{cases} 1 & \text{trt. A subject} \\ 0 & \text{trt. B subject} \end{cases}$$

$$\text{At each } t, \ \frac{\text{hazard for trt. A}}{\text{hazard for trt. B}} = e^{\beta} = \theta \text{ (say)}$$

θ is the **relative risk** (hazard ratio), $\beta = \ln(\theta)$

- How to estimate θ (or β)?
  How to test $H_0 : \theta = \theta_0$ (or $\beta = \beta_0$)?
- Standard analysis: Cox PH model (Cox, 1972)

Generalized Logrank (GLR) Method

Mehrotra and Roth (2001)

- Let $t_1 < t_2 < \ldots < t_k$ be the recorded failure times. At t_i:

Trt	Fail	Survive	Total
A	d_{iA}	$n_{iA} - d_{iA}$	n_{iA}
B	d_{iB}	$n_{iB} - d_{iB}$	n_{iB}
Total	d_i	$n_i - d_i$	n_i

If $d_{iB} \sim B(n_{iB}, p_i)$ then $d_{iA} \approx B(n_{iA}, \theta p_i)$; p_i is a nuisance parameter

- Without ties $(d_i = 1)$, for a given θ:

$$E(d_{iA} \mid d_i, n_{iA}, n_{iB}, \theta, p_i) \equiv E_{iA}(\theta, p_i) = \frac{n_{iA}\theta(1-p_i)}{n_{iA}\theta(1-p_i) + n_{iB}(1-\theta p_i)}$$

$$V(d_{iA} \mid d_i, n_{iA}, n_{iB}, \theta, p_i) \equiv V_{iA}(\theta, p_i) = \frac{n_{iA} n_{iB}\theta(1-p_i)(1-\theta p_i)}{[n_{iA}\theta(1-p_i) + n_{iB}(1-\theta p_i)]^2}$$

- **GLR estimator**: numerical solution of $GLR\left(\theta, \underline{\tilde{p}}(\theta)\right) = 0$

$$GLR\left(\theta, \underline{\tilde{p}}(\theta)\right) = \frac{\left\{\sum_{i=1}^{k}\left[d_{iA} - E_{iA}\left(\theta, \tilde{p}_{i,\theta}\right)\right]\right\}^2}{\sum_{i=1}^{k} V_{iA}\left(\theta, \tilde{p}_{i,\theta}\right)}$$

$$\tilde{p}_{i,\theta} = \frac{x_i - \sqrt{x_i^2 - 4n_i d_i \theta}}{2n_i\theta} = \text{unconditional m.l.e. of } p_i \mid \theta$$

$$x_i = \theta\left(n_{iA} + d_{iB}\right) + \left(n_{iB} + d_{iA}\right)$$

Note: $\tilde{p}_{i,\theta} \to 0$ as $n_i \to \infty$

- $GLR(\theta = 1, *)$ = logrank statistic (Mantel, 1966); it does not require estimation of the p_i's
- $GLR\left(\theta, \underline{p}(\theta) = 0\right) =$ score statistic from Cox (1972) model

GLR inference

- Reference distribution for $GLR[\theta_0, \underline{\tilde{p}}(\theta_0)]$?

  $GLR\left[\theta_0, \underline{\tilde{p}}(\theta_0)\right] \dot{\sim} F(1, k^*)$ under H_0

  where $k^* = \sum_{i=1}^{k} \min(d_i, n_i - d_i, n_{iA}, n_{iB})$

  $F(1, k^*) \to \chi^2_1$ as $k^* \to \infty$.

  $100(1-\alpha)\%$ confidence interval for θ

- $\theta^L_{GLR} = \inf_{\theta} \{\theta : GLR[\theta, \underline{\tilde{p}}(\theta)] \le F_\alpha(1, k^*)\}$

  $\theta^U_{GLR} = \sup_{\theta} \{\theta : GLR[\theta, \underline{\tilde{p}}(\theta)] \le F_\alpha(1, k^*)\}$

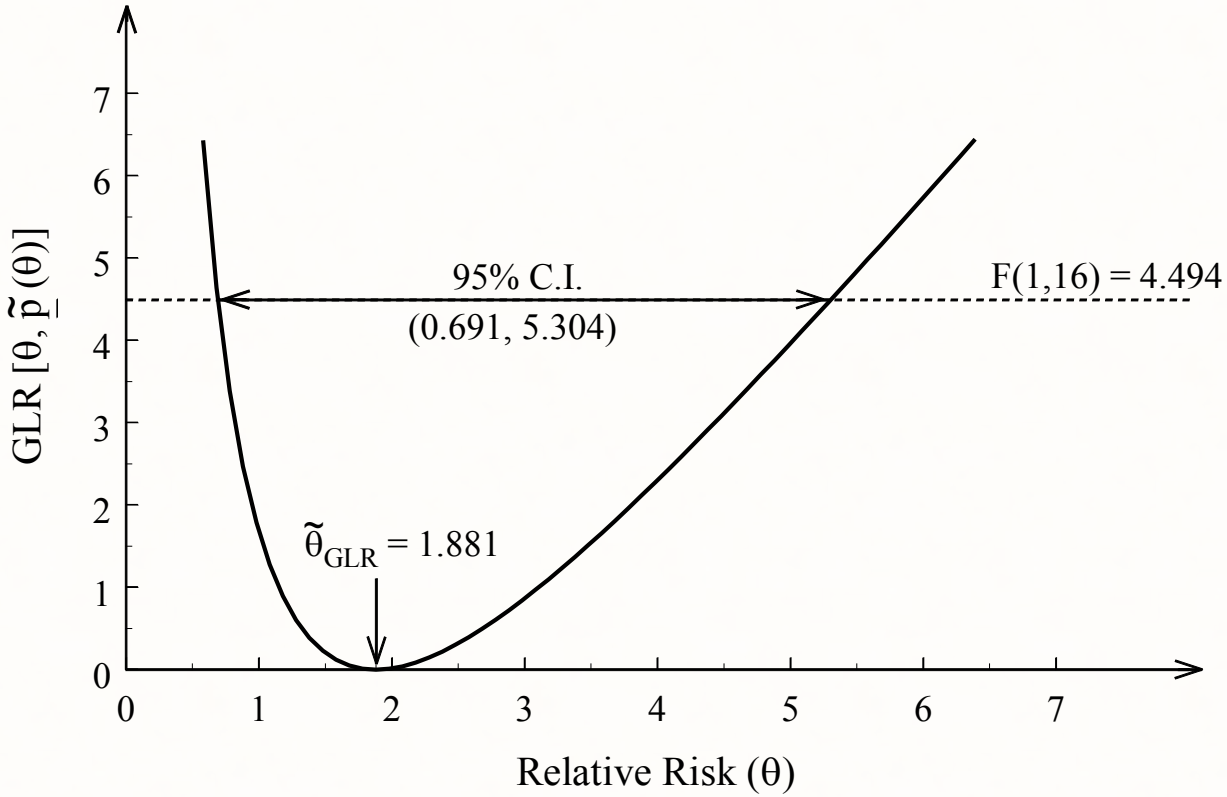
Example 1

Survival times (days) of 30 patients with cervical cancer

Data source: Parmar and Machin, 1995

Trt A (control): 90, 142, 150, 269, 291, 468+, 680, 837, 890+, 1037, 1090+, 1113+, 1153, 1297, 1429, 1577+

Trt B (new): 272, 362, 373, 383+, 519+, 563+, 650+, 827, 919+, 978+, 1100+, 1307, 1360+, 1476+



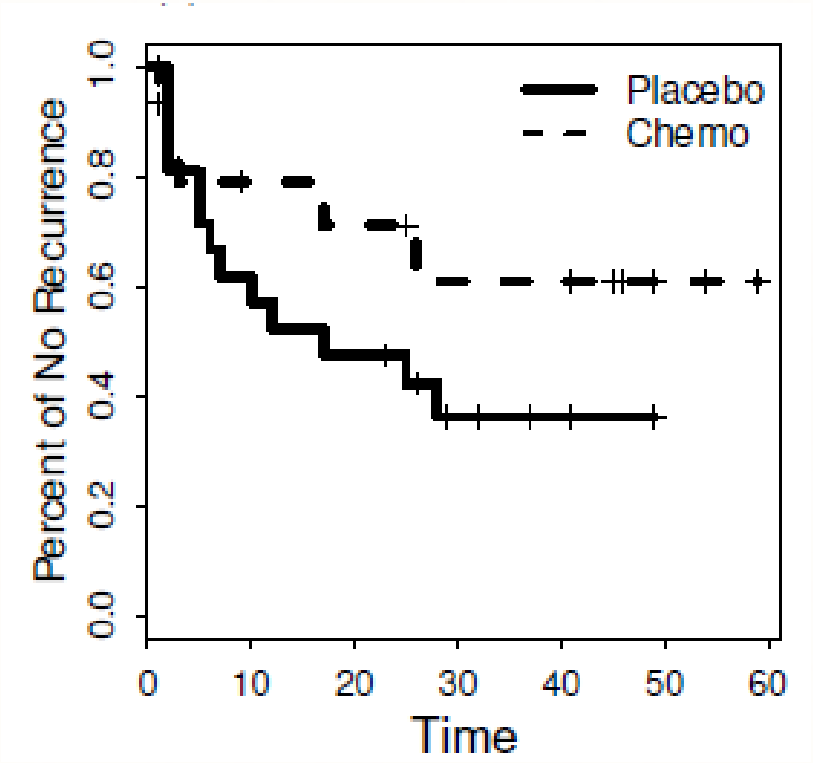
Cox model: $\tilde{\theta}_C = 2.00$, 95% C.I. = (0.69, 5.80)

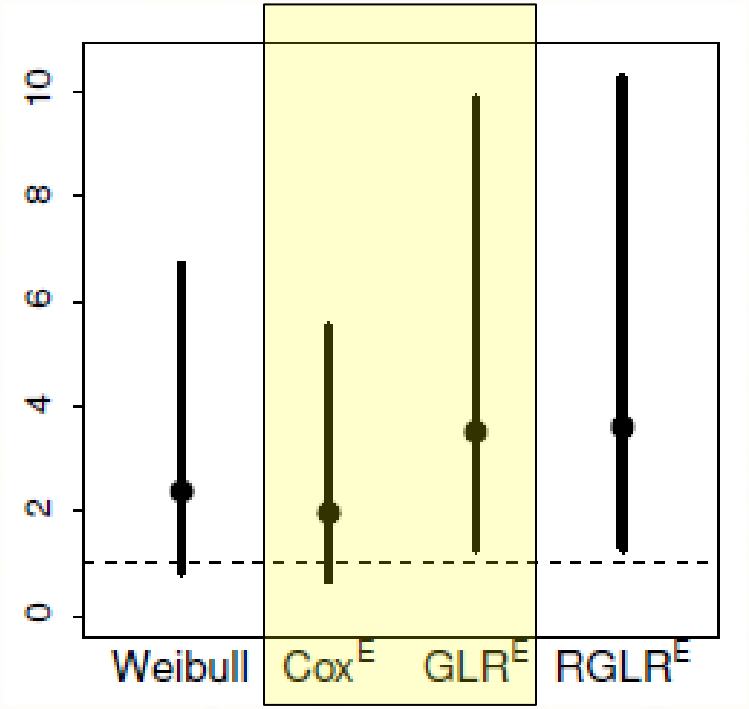
GLR method: $\tilde{\theta}_{GLR} = 1.88$, 95% C.I. = (0.69, 5.30)

Example 2

Survival times (days) of 37 patients with bladder cancer

Details: Xu et al (2017)





Estimated HR (95% CI)

Cox model: 1.96 (0.70, 5.51)

GLR method: 3.50 (1.27, 9.85)

Efron (E) approximation used for handling ties

Simulation Results

(Mehrotra and Roth 2001)

|%Bias| for $\beta \in [0.6, 1.2, 1.6]$

Cox 0.5% to 5.1%

GLR 0.4% to 3.8%

Mean Squared Error (MSE) and Relative Efficiency

		No censoring			**50% censoring**		
N/group	β	Cox	GLR	%RE	Cox	GLR	%RE
20	0.0	.121	.102	**119**	.211	.184	**114**
	1.6	.196	.160	**123**	.328	.261	**126**
40	0.0	.055	.050	**111**	.109	.101	**107**
	1.6	.089	.082	**109**	.150	.133	**113**
100	0.0	.021	.020	**105**	.039	.038	**103**
	1.6	.033	.032	**103**	.054	.051	**105**

%RE = $100 \times (MSE_{Cox})/(MSE_{GLR})$; 2000 replications

Conclusions: Topic #3

- Without ties: the GLR estimator is more efficient than the Cox estimator. (Mehrotra and Roth, 2001)

  With ties: the GLR^{KP} and GLR^{E} estimators are more efficient than the Kalbfleisch-Prentice and Efron estimators, respectively. (Mehrotra and Roth, 2011)

- GLR-based methods are useful for:
  (1) Small trials (e.g., cancer or rare disease)
  (2) Adaptive trials with early interim looks based on small samples when there is large uncertainty about the true hazard ratio.

- SAS code for GLR methods is available from the author.

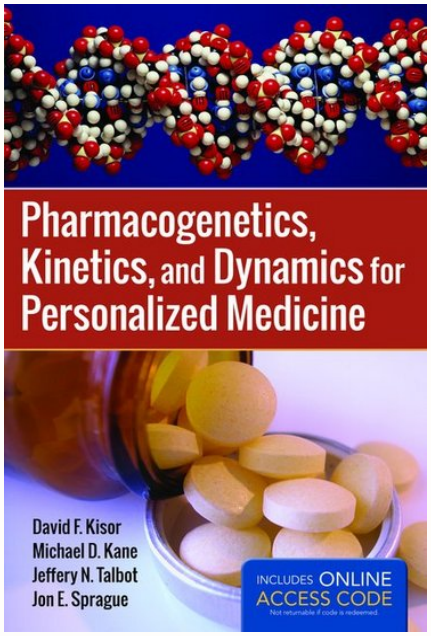
References: Topic #3

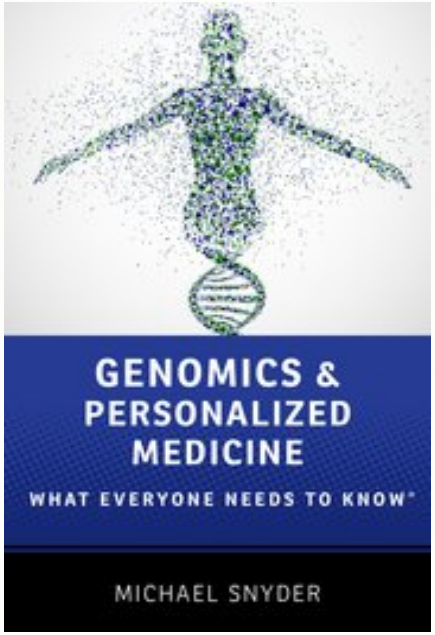
- Cox DR (1972). Journal of the Royal Statistical Society, B, 34, 187-220
- Efron B (1977). Journal of the American Statistical Association, 72, 557-565
- Kalbfleisch JD, Prentice RL (1973). Biometrika, 60, 267-278
- Mantel N (1966). Cancer Chemotherapy Reports, 50, 163-170
- Mehrotra DV, Roth AJ (2001). Statistics in Medicine, 20, 2099-2113
- Mehrotra DV, Roth AJ (2011). Statistics in Biopharmaceutical Research, 3, 456-462
- Parmar MKB, Machin D (1995). Survival Analysis: A Practical Approach, John Wiley and Sons: Chichester
- Xu R, Shaw PA, Mehrotra DV (2017). Statistics in Biopharmaceutical Research. https://www.tandfonline.com/doi/full/10.1080/19466315.2017.1369899

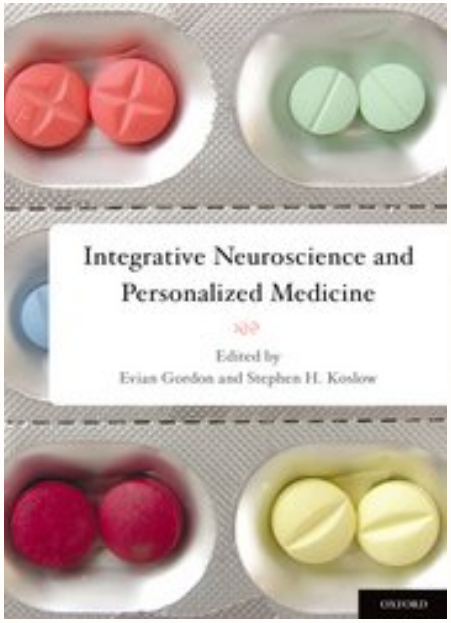
Topic #4

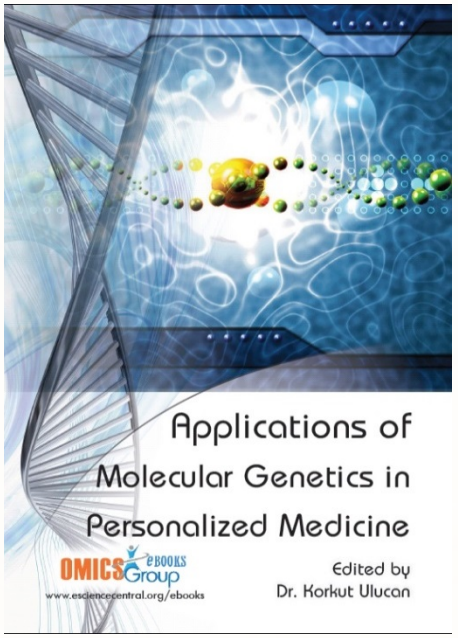
Pharmacogenomics in Phase 2-3 Trials

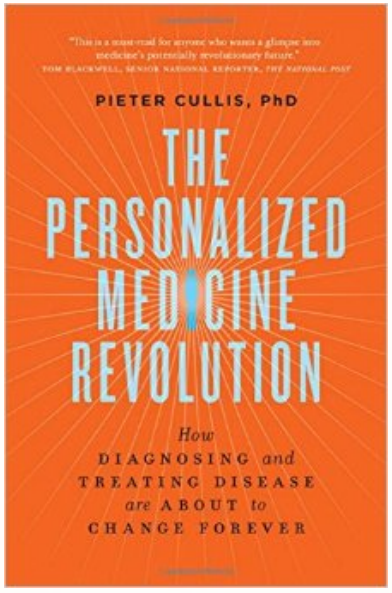
Precision/Personalized Medicine: A lot of buzz!

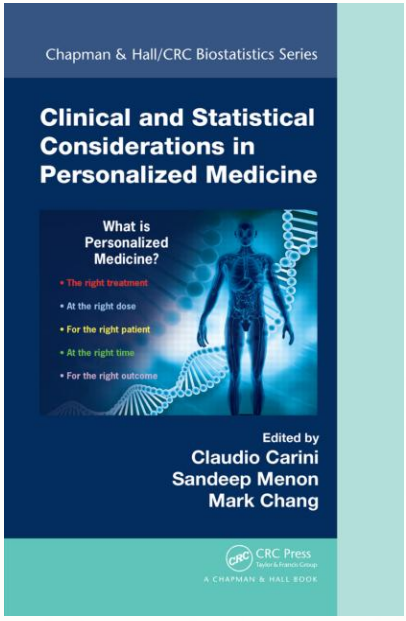


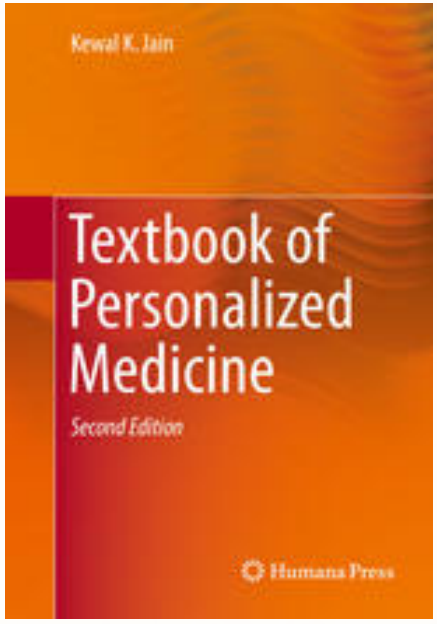




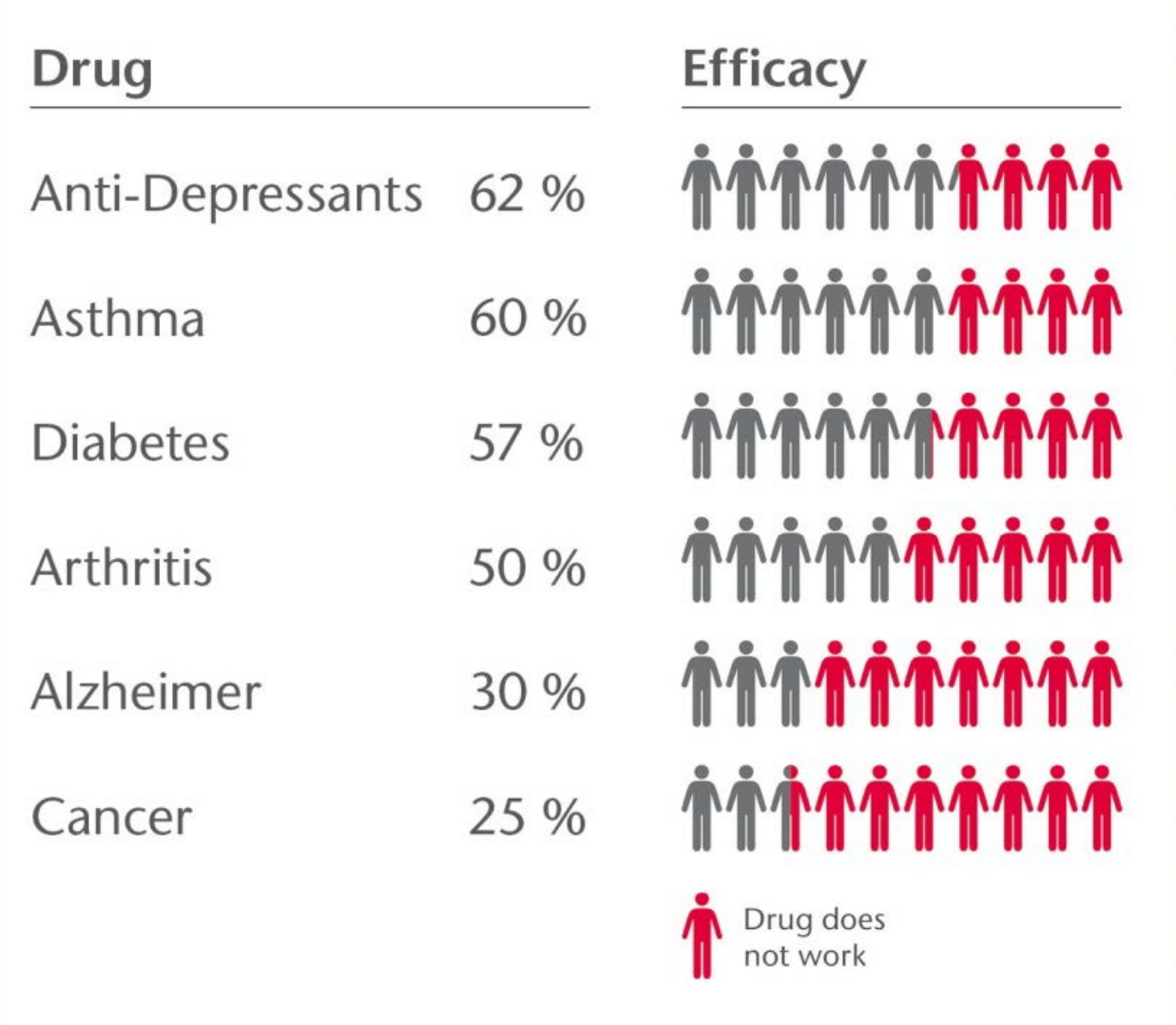








Precision/Personalized Medicine: Why?

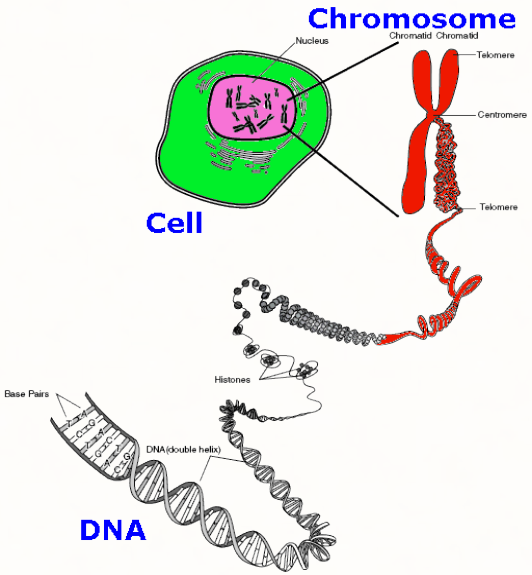


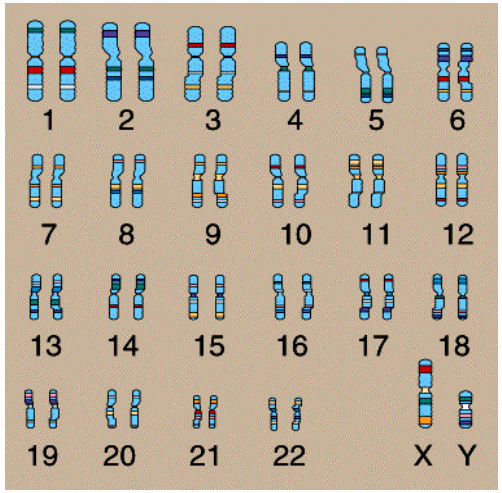
http://companion-diagnostics.com/personalized-healthcare/

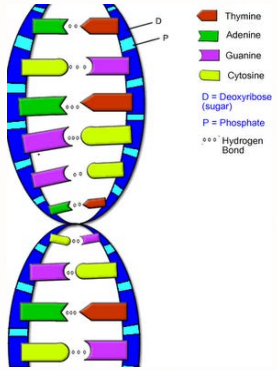
Key Messages of this Topic

- Discovery of genomic markers of **disease** usually requires sample sizes of 10,000+
- With a focus on personalized medicine, we show:
  1. How to discover actionable genomic markers of **drug response** with only ~ 300-500 subjects
  2. How to do a "winner's curse" adjustment when planning a confirmation study
- Applications: multi-arm dose-response (phase 2) and two-arm confirmatory (phase 3) trials

Single Nucleotide Polymorphism (SNP)







Nucleotide base pairs

A ⇔ T

G ⇔ C

Genome – "book of life"

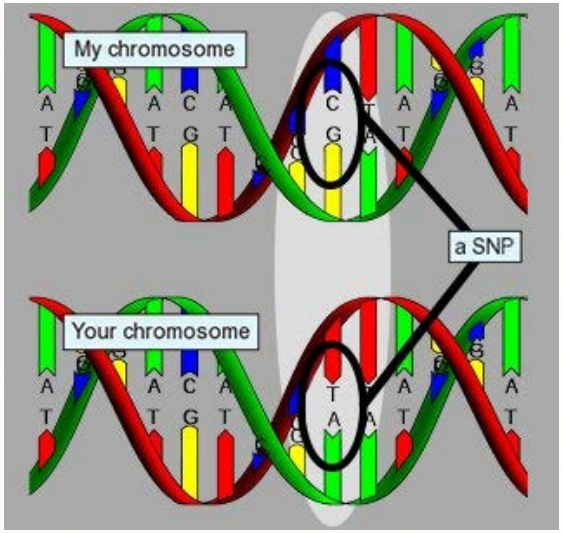
23 chapters called **chromosomes**

Each chapter has 1000s of stories called **genes**

Each story has paragraphs called **exons** (w/**intron** gaps)

Each paragraph has words called **codons**

Each word is written in letters called **bases**



Single Nucleotide Polymorphism (SNP)

Matt Ridley's "Genome" (1999) http://images.google.com

Statistical Methods

- **Context**: Genomewide association study (GWAS) within a two-arm phase III randomized clinical trial
- Treatment A [old] vs. Treatment B [new]
- Y = subject-level response to treatment (yes/no)

  p = E(Y) = true proportion of responders

  T = treatment (0=A, 1=B)

  G = SNP genotype subgroup (0=SNP-, 1=SNP+)
- **Goal**: to identify SNPs for which there is a treatment by genotype (TxG) interaction

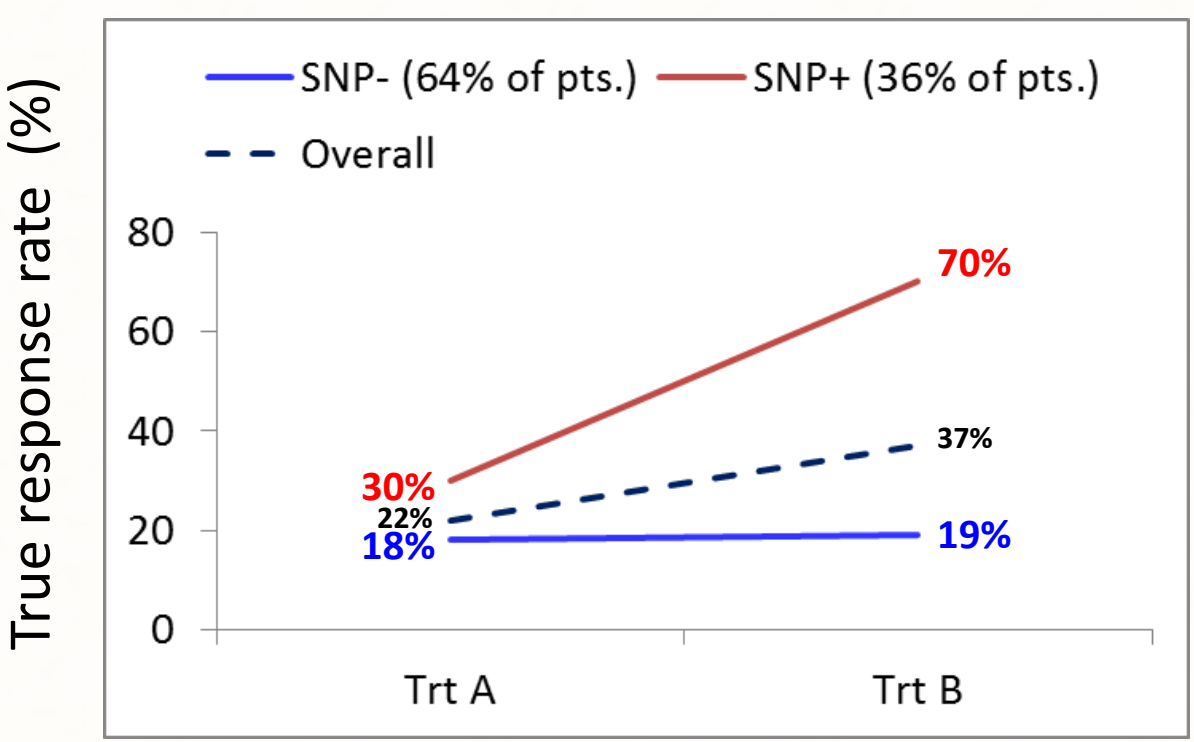
Statistical Methods (continued)

- Analysis model: $\text{logit}(p) = \beta_0 + \beta_T T + \beta_G G + \beta_{TG} TxG$
  (other covariates can be added)
- **Method 1 (traditional)**
  - Test H_{null}: $\beta_{TG}=0$
  - 1 df likelihood ratio test

  Method 2 (proposed)
  - Test H_{null}: $\beta_G=\beta_{TG}=0$
  - 2 df likelihood ratio test
  - Why? TxG interactions rarely exist in the absence of G main effects. This is a **much more powerful** way to discover SNPs with TxG interactions!

Sample Size Comparison: Method 1 vs. Method 2

- Assumption (at design stage): p_A=22% (old trt), p_B=37% (new trt)
- For 90% power: N=190 per trt grp
- **Truth**: B is clinically better than A in **SNP+** subjects only (specific SNP)
- What is the power to discover this SNP in a GWAS?



Power to discover SNP (at α=5E-8)

Method 1: <1%
Method 2: 95%

↓

Trial has good power to discover this SNP only with Method 2

"Significant" SNPs: Winner's Curse

- **Discovery phase**: for any "significant" SNP (2 df $p < 5E\text{-}8$), the parameter estimates will be biased (winner's curse)

  Bias is worse when the SNP discovery power is small!

- **Confirmation phase**: use of the naïve (i.e., unadjusted) parameter estimates will result in under-powered study

- We have developed a **conditional maximum likelihood** approach to adjust for selection bias when planning a confirmation study (next slide)

Winner's Curse Adjustment

Conditional Maximum Likelihood Estimator (MLE)

$$\hat{\beta}_{MLE} = \text{argmax}_{\beta} L(\beta)$$

$$L(\boldsymbol{\beta}) = f_{Y|\Lambda \geq c}(Y;\boldsymbol{\beta}) = \frac{\prod_{i=1}^{N} \mu_i(\boldsymbol{\beta})^{Y_i}(1-\mu_i(\boldsymbol{\beta}))^{(1-Y_i)}}{\int_c^{\infty} f_{\chi^2}(t;s,\phi)dt}\mathbf{1}(\Lambda \geq c)$$

$$\boldsymbol{\beta} = (\beta_0, \beta_T, \beta_G, \beta_{TG}) \quad \mathbf{X}_i = (1, T_i, G_i, T_i \times G_i) \quad \mu_i(\boldsymbol{\beta}) = \frac{\exp(\mathbf{X_i}\beta)}{1+\exp(\mathbf{X_i}\beta)}$$

$$\phi \approx 2\sum_{i=1}^{N}\left\{\mu_i(\boldsymbol{\beta})(\mathbf{X}_i\boldsymbol{\beta} - \mathbf{X}_i\boldsymbol{\beta}^*) - \log\left[\frac{1+\exp(\mathbf{X_i}\boldsymbol{\beta})}{1+\exp(\mathbf{X_i}\boldsymbol{\beta}^*)}\right]\right\}$$

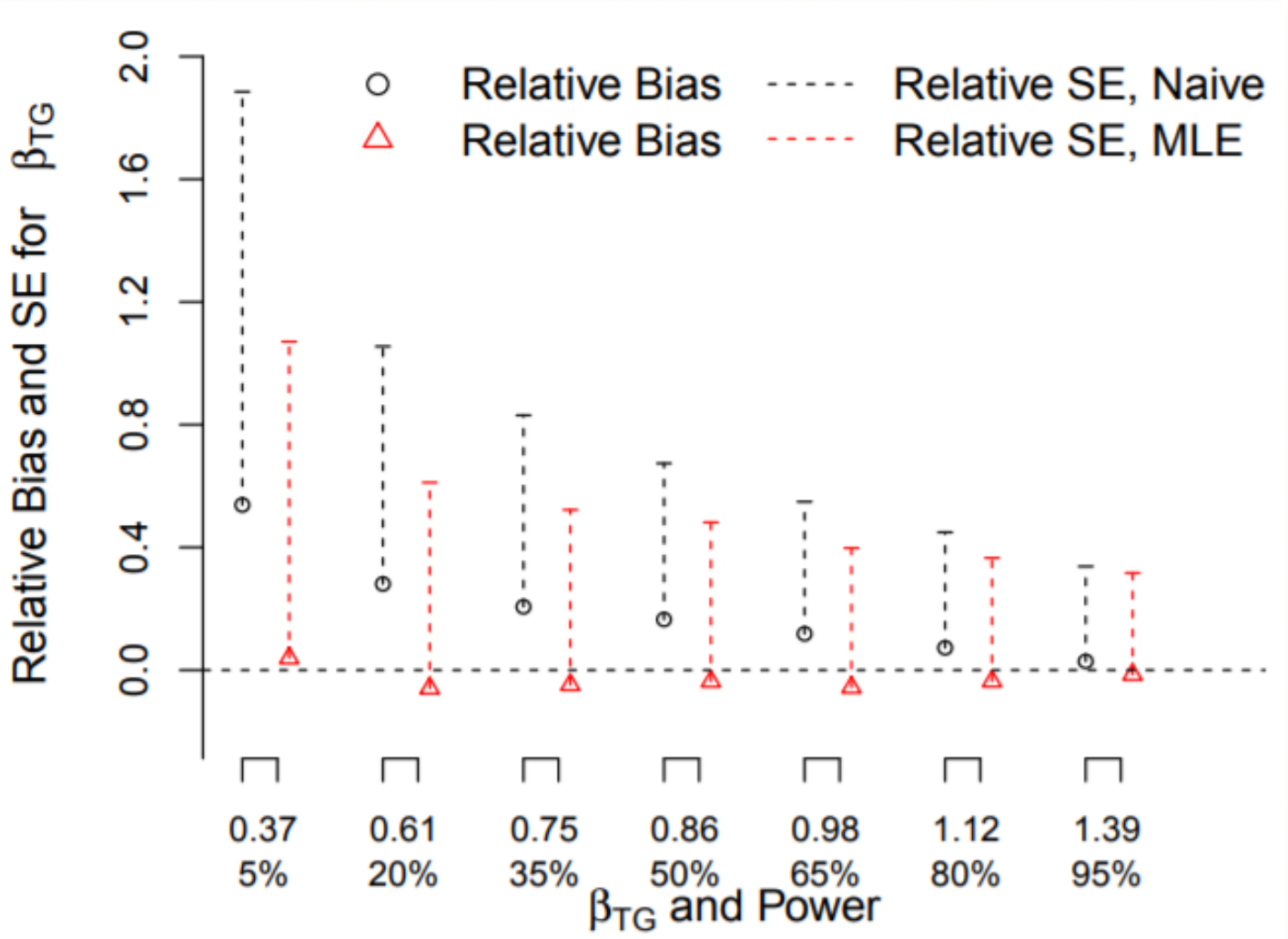
$$\boldsymbol{\beta}^* = (\beta_0^*, \beta_T^*, 0, 0)^T$$

$$0 = \sum_{i=1}^{N}[\mu_i(\boldsymbol{\beta}) - \mu_i((\beta_0^*, \beta_T^*, 0, 0)^T)]\mathbf{X}_i$$

Details in manuscript under preparation

Naïve Estimator vs. Conditional MLE of β_{TG}

Simulation Results: Bias and Standard Error



N=500 (250/trt); 1000 simulations; other details in the manuscript

Example: Treatment of Hepatitis C Infection

Double-blind, randomized control trial (Trt A vs. Trt B)

Efficacy Analysis

Sustained Viral Response (%)

Trt A Placebo*	Trt B New Drug*
46% (71/153)	78% (111/143)

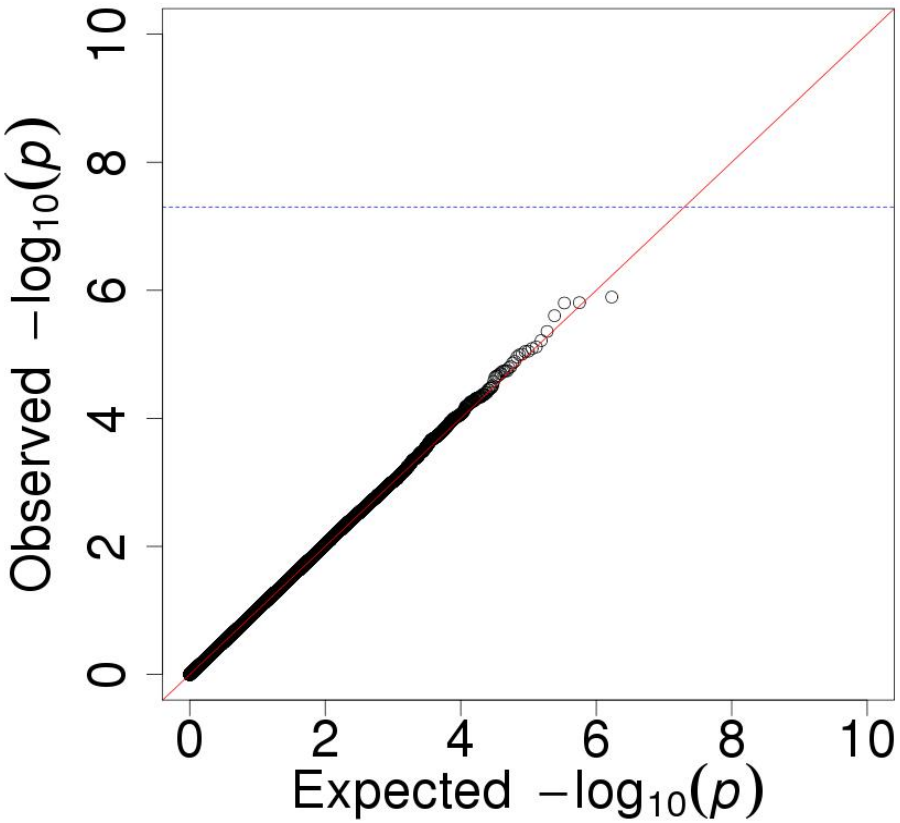
* added to standard of care (SOC)

- Overall, trt B has much better efficacy that trt A ($p < 0.0001$)
- Should the analysis stop here? NO

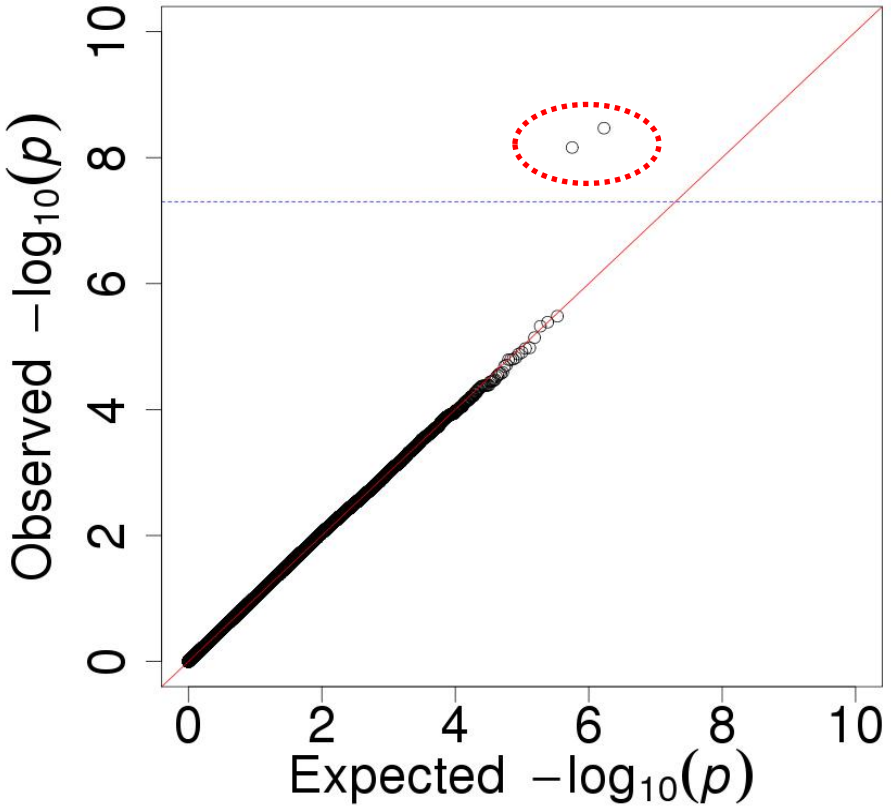
Example (continued): GWAS Results

GWAS: Q-Q Plots
(1 p-value for each SNP)

Method 1



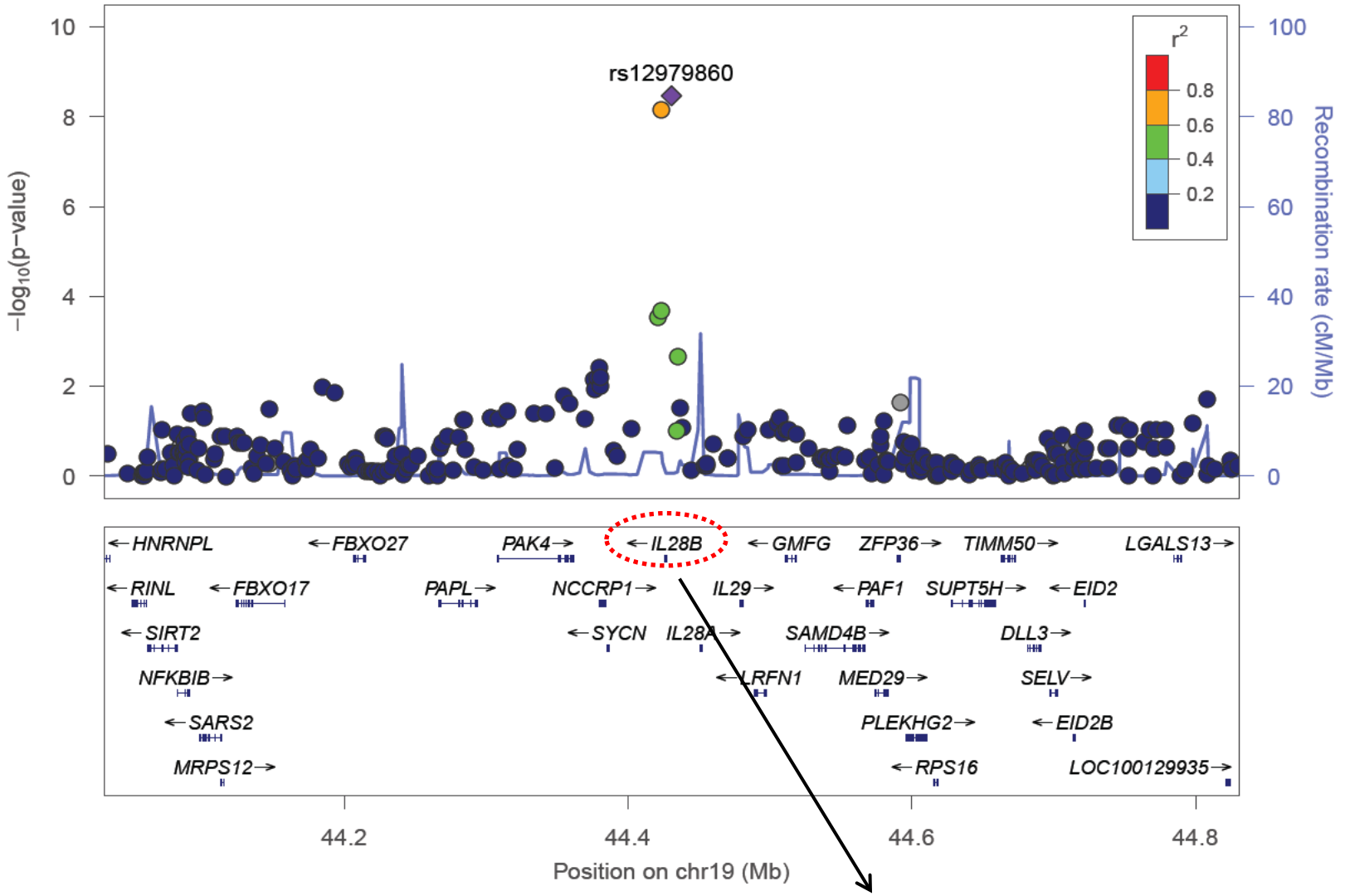
Method 2



Analysis model: logit(p) = $\beta_0 + \beta_T T + \beta_G G + \beta_{TG} TxG$

Method 1 (traditional) H_{null}: $\beta_{TG}=0$, Method 2 (proposed) H_{null}: $\beta_G=\beta_{TG}=0$

Example (continued): Region Plot



Discovered SNP (rs12979860) sits on **IL28B** gene

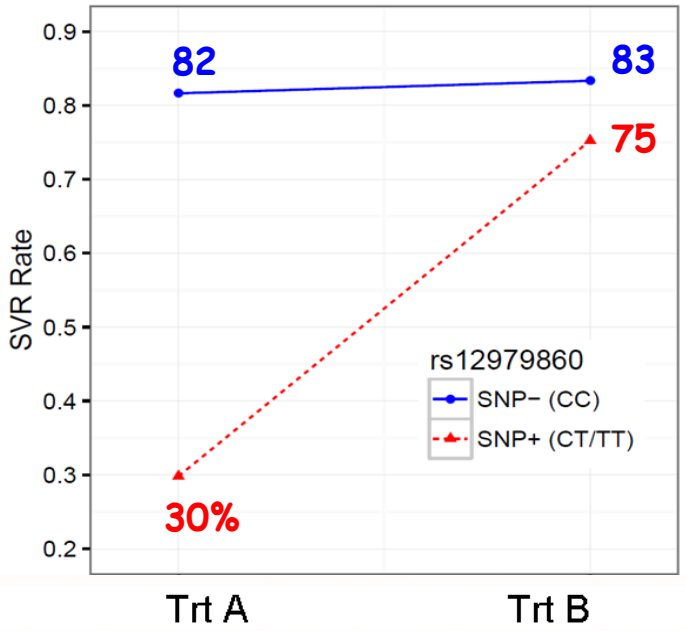
Example (continued): GWAS identifies important subgroups

Trt B has notably better SVR than Trt A in **SNP+** patients only

Sustained Viral Response (%)

SNP-based subgroup	Trt A Placebo*	Trt B New Drug*
SNP- **~ 30% of pts.**	82% (40/49)	83% (35/42)
SNP+ **~ 70% of pts.**	30% (31/104)	75% (76/101)
Overall	46%	78%

* both A and B were add-ons to standard of care



P-values for SNP (rs12979860)

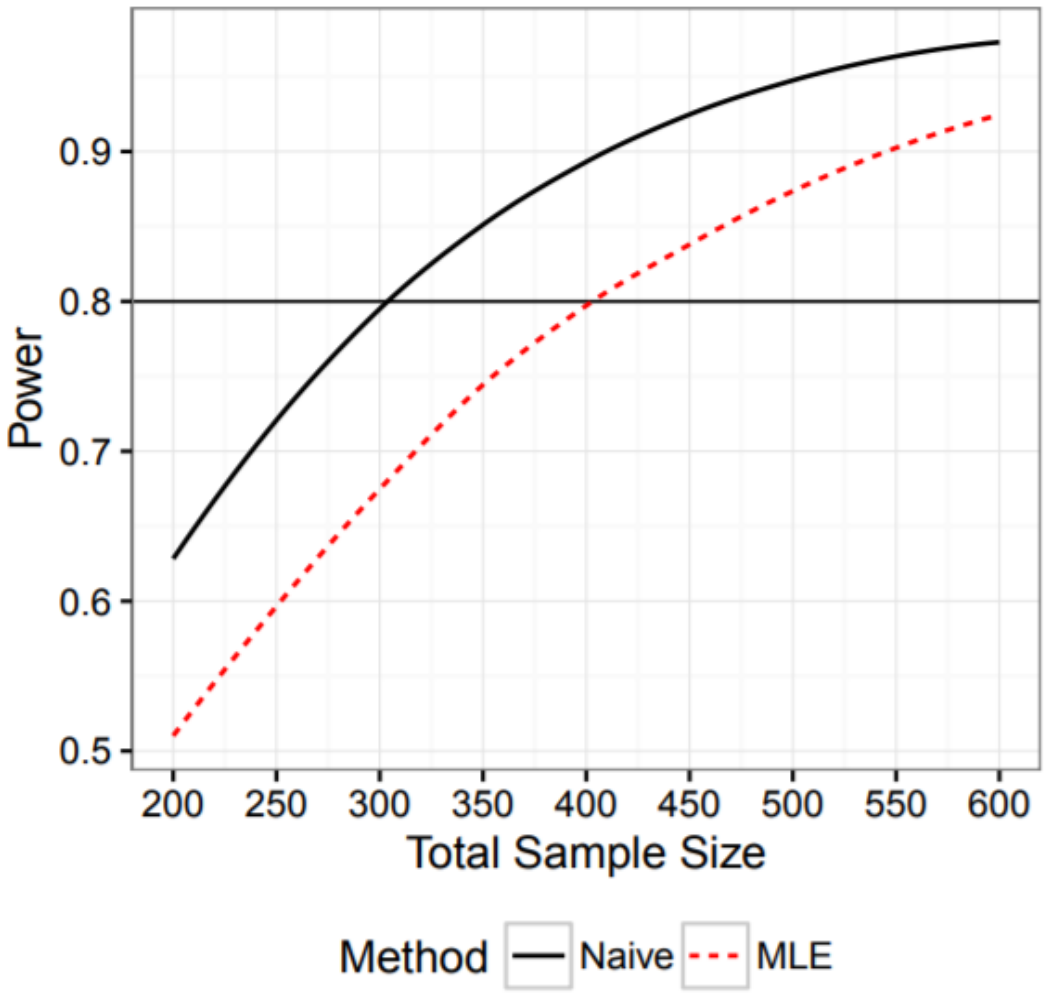
p-value for β_{TG}	p-value of Joint Test (β_G and β_{TG})
6.08E-03	3.41E-09
Method 1	**Method 2**

Parameter Estimates

Method	β_T	β_G	β_{TG}
Naive	0.16	-2.38	1.83
MLE	0.38	-1.97	1.50

Example (continued): N for Confirmation Study?

SNP = rs12979860



Total N for 80% Power*
(to confirm TxG interaction)

Naïve: N = 300**
MLE: N = 400

* assuming β_{TG} = point estimate

** actual power = 67% if $\beta_{TG} = 1.50$

Winner's curse adjustment

Method	β_T	β_G	β_{TG}
Naive	0.16	-2.38	1.83
MLE	0.38	-1.97	1.50

GWAS: Dose-Response Trials

Statistical Methodology

- p = true proportion of treatment "responders"

  T = treatment dose level (coded 0, 1, 2, …)

  G = genotype subgroup (coded 0,1,2)

  Traditional Method

  $\text{logit}(p) = \beta_0 + \beta_1 T + \beta_2 G + \beta_3 TxG$

  Test H_{TG}: $\beta_3=0$ (same G effect for all doses)

  Proposed Method*

  $\text{logit}(p) = \beta_0 + \beta_1 T + \beta_2 G + \beta_3 TxG$

  Test $H_{G,TG}$: $\beta_2=\beta_3=0$ (no G effect for any dose)

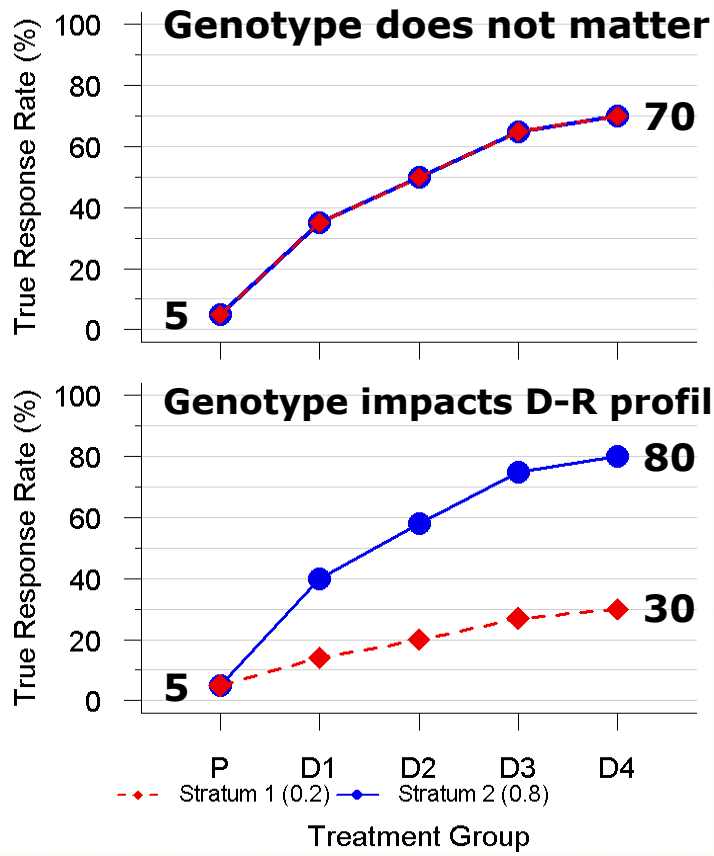
  Covariates can be added to both logistic regression models

  (* extension of preceding slides to dose-response setting)

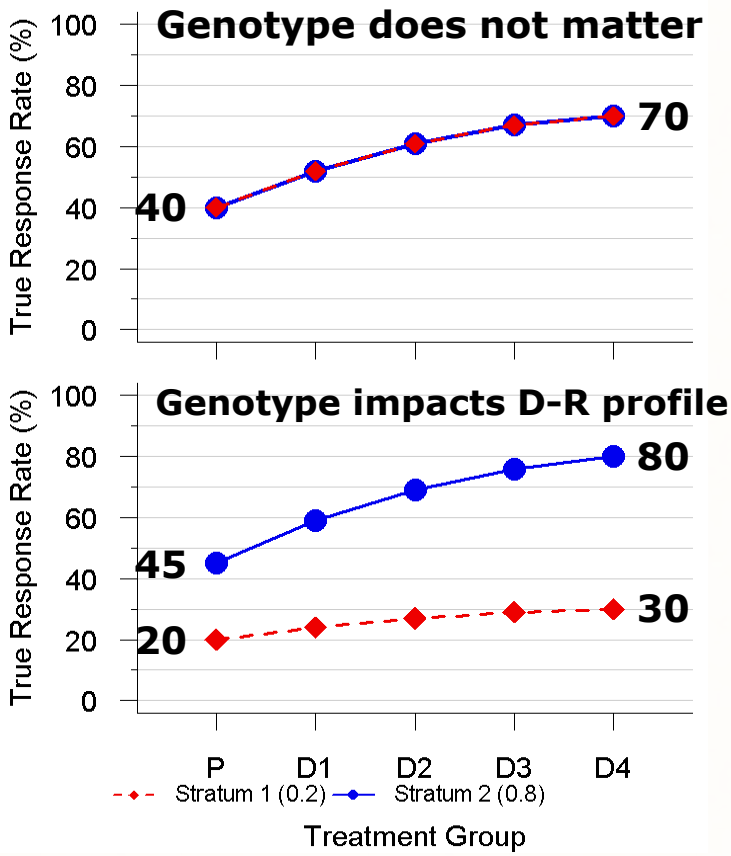
GWAS Power Boost for D-R Trials: Simulations

Proposed method delivers adequate power with a relatively small N

Scenario 1: Low Placebo Response



Scenario 2: High Placebo Response



N for 80% power to detect SNP effect

Traditional Method: 820/dose group
Proposed Method: 80/dose group

10-fold reduction in sample size!

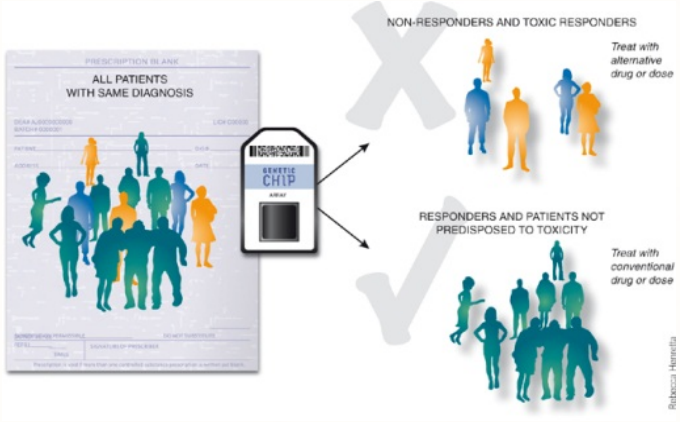
N for 80% power to detect SNP effect

Traditional Method: 1600/dose group
Proposed Method: 70/dose group

20-fold reduction in sample size!

Conclusions: Topic #4

- Our proposed methods enable the discovery of genotype-phenotype associations with relatively small sample sizes in phase 2-3 clinical trials
- If any discovered SNP is deemed "actionable" based on patient/prescriber/payer input, think about:
  - Strategy for <u>confirmation</u> of the discovery
  - Strategy for companion/complementary diagnostic



References: Topic #4

- Dai JY, et al (2012). American Journal of Epidemiology, 176, 164-173
- Kraft P, et al. (2007). Human Heredity, 63, 111-119.
- Lunceford JK, et al (2014). Statistics in Biopharmaceutical Research, 6, 137-143.
- Mehrotra DV, Guan Q, Guo Z. A Statistical Boost for Embedding Pharmacogenomics in Clinical Drug Development. To be submitted to Nature Genetics.

Topic #5

Stratified Trials: Binary or Time-to-Event Endpoints

Why Stratify?

Illustrative Example

Percentage of "responders" to treatment A and B

	Treatment A (new)	Treatment B (old)	A - B
	50% (50/100)	52% (52/100)	-2%
Biomarker+	88% (35/40)	77% (46/60)	+11%
Biomarker-	25% (15/60)	15% (6/40)	+10%
Pooled	50% (50/100)	52% (52/100)	-2%

Pooled = ignoring biomarker status

Failure to stratify on **prognostic factor(s)** can yield misleading and/or inefficient analyses

Prognostic vs. Non-Prognostic Factors

- **Prognostic** factor: likely to influence patient response to treatment in a systematic manner

  Examples: age, sex, disease severity, biomarker

- **Non-prognostic** factor: unlikely to influence patient response to treatment in a systematic manner

  Example: study center

  Note: okay to stratify randomization by center but not necessary to "adjust" for center in the analysis

Prognostic Factor Examples: Merck Trials

Prognostic Factor	Levels	% Responders	Response definition
Baseline Ad5 titer	≤ 200 > 200	80% 25%	Response to HIV Ad5-gag vaccine (ELISPOT assay)
CD4 count (cells/mm^3)	≥ 50 < 50	44% 21%	vRNA < 400 c/mL at 48 wks (placebo+SOC arm)
IL28B SNP (rs12979860)	SNP- SNP+	73% 28%	Sustained Virological Response (peg + riba arm)
Prior CDI infection?	Yes No	47% 27%	Recurrence of CDI (placebo arm)

In every example above: ≥ **20% absolute difference** in response rate between the two subgroups

Binary Endpoint

Superiority trials

Motivating Example

- A randomized clinical trial is being designed to compare a new treatment (A) to an old treatment (B) based on a binary endpoint (responder/non-responder).

- **Clinician**: "Historically, 75% of patients have responded to the old treatment and we believe 85% will respond to the new treatment. What **sample size** do we need to detect this treatment difference (75% → 85%)?"

- **Statistician**: "For 80% power at 1-tailed alpha=.025, we need n=250 per treatment group". [Total N=500]

$$n = \frac{\left\{z_\alpha \sqrt{2\bar{p}(1-\bar{p})} + z_\beta \sqrt{p_A(1-p_A) + p_B(1-p_B)}\right\}^2}{(p_A - p_B)^2}$$

$$z_\alpha = 1.96, z_\beta = 0.842, p_A = .85, p_B = .75, \bar{p} = (.85 + .75)/2 = .80$$

Motivating Example (continued)

- **Clinician**: "A total N of 500 will cost us $6 million, but adequate power is important. By the way, we expect 65% of males and 85% of females to respond to the old treatment, and about half the population is male."
- **Statistician**: "We should stratify randomization by sex and pre-specify the stratified **Miettinen and Nurminen (MN) method** in the SAP. Can we assume the same treatment difference in response rates for each sex?"
- **Clinician**: "I don't know … let's get input from four independent subject matter experts."

Motivating Example (continued)

- The SMEs have slightly different opinions on the expected "delta" (treatment difference) for each sex.

	Males δ_M	Females δ_F	Overall $\delta = 0.5\delta_M + 0.5\delta_F$
Expert 1	**10%** (65 → 75)	**10%** (85 → 95)	**10%** (75 → 85)
Expert 2	**10%** (65 → 75)	**9%** (85 → 94)	**9.5%** (75 → 84.5)
Expert 3	**9%** (65 → 74)	**10%** (85 → 95)	**9.5%** (75 → 84.5)
Expert 4	**7%** (65 → 72)	**11%** (85 → 96)	**9.0%** (75 → 84)

- **Clinician**: "Does our trial have adequate power across all four scenarios if the stratified **MN method** is used?"

Motivating Example (continued)

- **Statistician**: "I simulated the four scenarios … we should switch from the MN to the **MR method**!"

Expert	Males δ_M	Females δ_F	Overall δ	Power (%)		
				Unstratified	MN	**MR**
Expert 1	10%	10%	10%	80	83	**87**
Expert 2	10%	9%	9.5%	75	78	**82**
Expert 3	9%	10%	9.5%	75	78	**84**
Expert 4	7%	11%	9%	70	74	**82**

100,000 simulations; all methods control the type I error rate (not shown)

- **Clinician**: "I'm convinced! But how do the MN and MR methods differ? Have both been used in real trials?"
- **Statistician**: "Both methods have been reported in JAMA, NEJM and Lancet papers. Let me explain how they differ."

Stratified Test: MN vs. MR

MN = Miettinen and Nurminen (1985), MR=Mehrotra and Railkar (2000)

$$H_{null}: \delta = 0 \quad \text{vs.} \quad H_{alt}: \delta > 0$$

$$Z_{cal} = \frac{\sum_i w_i \hat{\delta}_i - cc}{\sqrt{\sum_i a_i w_i^2 V(\hat{\delta}_i)}}$$

$$p - value = P(Z > Z_{cal})$$

$\hat{p}_{ij}$ = observed success proportion for stratum i, trt j

$\hat{\delta}_i = \hat{p}_{iA} - \hat{p}_{iB}$, $\hat{\delta} = \sum_i w_i \hat{\delta}_i$

w_i = weight for stratum i

a_i = finite sample term for stratum i

cc = continuity correction

MN and MR tests use different $V(\hat{\delta}_i)$, w_i, a_i, and cc

Stratified Test: MN vs. MR (continued)

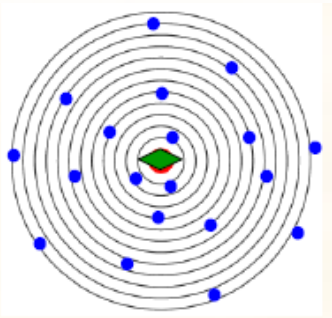
	Stratified MN test	Stratified MR test
$V(\hat{\delta}_i)$	$\left(\frac{\overline{p}_i\overline{q}_i}{n_{iA}}+\frac{\overline{p}_i\overline{q}_i}{n_{iB}}\right)\equiv V_{i,null}$	$\left(\frac{\hat{p}_{iA}\hat{q}_{iA}}{n_{iA}}+\frac{\hat{p}_{iB}\hat{q}_{iB}}{n_{iB}}\right)\equiv V_{i,obs}$
w_i	CMH weights	MR weights
a_i	$(n_{iA}+n_{iB})/(n_{iA}+n_{iB}-1)$	1
cc	0	$\frac{3}{16}\left(\sum_i \frac{n_{iA}n_{iB}}{n_{iA}+n_{iB}}\right)^{-1}$

$\overline{p}_i=(n_{iA}\hat{p}_{iA}+n_{iB}\hat{p}_{iB})/(n_{iA}+n_{iB});\ \overline{q}_i=1-\overline{p}_i$

Note: $V_{i,obs}\le V_{i,null}$ for 1:1 randomization.

Stratified Test: MN vs. MR (continued)

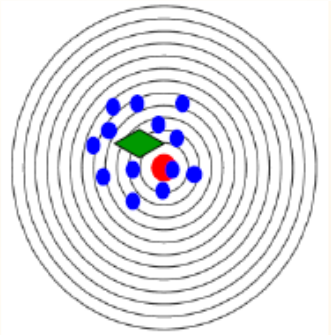
- **Cochran-Mantel-Haenszel (CMH)** weights



$$w_i^{CMH} = \frac{n_{iA}n_{iB}/(n_{iA}+n_{iB})}{\sum_i n_{iA}n_{iB}/(n_{iA}+n_{iB})}$$

Goal: minimize $\left|E(\hat{\delta}-\delta)\right|$ {bias}

- **Minimum Risk (MR)** weights



For two strata trial:

$$w_1^{MR} = \frac{V_{1,obs}^{-1} + \frac{(n_{1A}+n_{1B})}{\sum_i (n_{iA}+n_{iB})}(\hat{\delta}_1-\hat{\delta}_2)^2 V_{1,obs}^{-1}V_{2,obs}^{-1}}{V_{1,obs}^{-1}+V_{2,obs}^{-1}+(\hat{\delta}_1-\hat{\delta}_2)^2 V_{1,obs}^{-1}V_{2,obs}^{-1}}, \quad w_2^{MR} = 1 - w_1^{MR}$$

Goal: minimize $E(\hat{\delta}-\delta)^2$ {mean squared error}

Motivating Example: Results

$$H_{null}: \delta = 0 \quad \text{vs.} \quad H_{alt}: \delta > 0$$

δ = true overall treatment difference (A-B)

Observed percentage of responders				**Stratum Weights**	
Stratum	New Trt [A]	Old Trt [B]	A – B	MN method	MR method
Males	70.3% (104/148)	68.8% (99/144)	3.4%	.59	.51
Females	95.1% (97/102)	83.3% (85/102)	11.8%	.41	.49

Method	Estimated δ	SD of estimated δ	95% CI for δ	1-tailed p-value
MN	6.8%	3.7%	(-0.4,13.9)	.032
MR	7.5%	3.5%	(0.7,14.3)	.017*

* statistically significant (1-tailed p < .025)

Binary Endpoint

Non-inferiority trials

Hypothesis Test for Non-Inferiority

$$H_{null} : \delta \le -\delta_0 \text{ (inferiority)}$$

$$H_{alt} : \delta > -\delta_0 \text{ (non-inferiority)}$$

$\delta_0 (> 0)$ is the non-inferiority margin

$$Z_{cal} = \frac{\left(\sum_i w_i \hat{\delta}_i + \delta_0 \right) - cc}{\sqrt{\sum_i a_i w_i^2 V(\hat{\delta}_i)}}$$

$\hat{p}_{ij}$ = observed success proportion for stratum i, trt j

$\hat{\delta}_i = \hat{p}_{iA} - \hat{p}_{iB}, \hat{\delta} = \sum_i w_i \hat{\delta}_i$

w_i = weight for stratum i

a_i = finite sample term for stratum i

cc = continuity correction

Choice of Variance

- **Null variance** [Miettinen & Nurminen, 1985] (MN)

$$V\left(\hat{\delta}_i\right) = \frac{\tilde{p}_{iA}\left(1-\tilde{p}_{iA}\right)}{n_{iA}} + \frac{\tilde{p}_{iB}\left(1-\tilde{p}_{iB}\right)}{n_{iB}} \equiv \tilde{V}_i$$

$\tilde{p}_{ij}$ = m.l.e. of p_{ij} under the restriction $p_{iA} - p_{iB} = -\delta_0$

Note: see also Farrington & Manning (1990)

- **Observed (OBS) variance**

$$V\left(\hat{\delta}_i\right) = \frac{\hat{p}_{iA}\left(1-\hat{p}_{iA}\right)}{n_{iA}} + \frac{\hat{p}_{iB}\left(1-\hat{p}_{iB}\right)}{n_{iB}} \equiv \hat{V}_i$$

- With 1:1 randomization, $\hat{V}_i$ is always $\leq \tilde{V}_i$ for superiority trials, and often (but not always) for non-inferiority trials.

Choice of Variance (continued)

Null Variance

- Formula for $\tilde{p}_{ij}$ from Farrington & Manning (1990)

$$\tilde{p}_{iA} = 2u_i \cos(w_i) - b_i (3a_i)^{-1}, \tilde{p}_{iB} = \tilde{p}_{iA} + \delta_0$$

where

$$a_i = 1 + \theta_i, \theta_i = n_{iB} n_{iA}^{-1}$$

$$b_i = -[1 + \theta_i + \hat{p}_{iA} + \theta_i \hat{p}_{iB} - \delta_0(\theta_i + 2)]$$

$$c_i = \delta_0^2 - \delta_0(2\hat{p}_{iA} + \theta_i + 1) + \hat{p}_{iA} + \theta_i \hat{p}_{iB}$$

$$d_i = \hat{p}_{iA}\delta_0(1 - \delta_0)$$

$$u_i = \text{sgn}(v_i)\sqrt{b_i^2 (9a_i^2)^{-1} - c_i (3a_i)^{-1}}$$

$$v_i = b_i^3 (3a_i)^{-3} - b_i c_i (6a_i^2)^{-1} + d(2a_i)^{-1}$$

$$w_i = (3)^{-1}[\pi + \cos^{-1}(v_i u_i^{-3})]$$

Motivating NI Example

- Trial objective: assess **non-inferiority** of a new treatment (A) to the old treatment (B) based on a binary endpoint.

  New treatment is expected to retain the efficacy of the old treatment but have a better safety profile

- **80%** of patients respond to the old treatment.

- The non-inferiority margin (Δ) is set at **10%**

$$H_{null} : \delta \leq -10\% \quad \text{vs.} \quad H_{alt} : \delta > -10\%$$

  δ is the true overall treatment difference

- Sample size required = **340 per treatment group** for 90% power, 1-tailed α=.025 (unstratified: Farrington & Manning, 1990)

  Assumptions: (i) δ=0, (ii) p_A=p_B=80%

Motivating NI Example (continued)

- Additional information available at the design stage:
  1. About half the patient population is $\geq$ 65 years
  2. Age is a prognostic factor for the old treatment

     Patients < 65 years: ~ 90% respond

     Patients $\geq$ 65 years: ~ 70% respond

- **Stratified randomization** (age < 65/$\geq$ 65 yrs) is planned.

- The "traditional" method of analysis for NI trials is the stratified **Miettinen and Nurminen (MN) method**.

- Statistician and clinician design a **simulation** study to check whether the trial is adequately powered based on the planned analysis using the stratified MN method.

Motivating NI Example (continued)

Simulated % responders and "deltas" (= $p_A - p_B$)

	Age < 65 yrs δ_1 (p_A,p_B)	Age ≥ 65 yrs δ_2 (p_A,p_B)	**Overall** $\delta = 0.5\delta_1+0.5\delta_2$ (p_A,p_B)
Scenario 1	**0%** (90%, 90%)	**0%** (70%, 70%)	**0%** (80%, 80%)
Scenario 2	**-1%** (89%, 90%)	**-1%** (69%, 70%)	**-1%** (79%, 80%)
Scenario 3	**-1%** (89%, 90%)	**-2%** (68%, 70%)	**-1.5%** (78.5%, 80%)
Scenario 4	**-2%** (88%, 90%)	**-1%** (69%, 70%)	**-1.5%** (78.5%, 80%)

Motivating NI Example (continued)

$$H_{null} : \delta \leq -10\% \quad \text{vs.} \quad H_{alt} : \delta > -10\%$$

Simulation Results

N=340/group (50% in each stratum), target power = 90%

	< 65 yrs δ_1	≥ 65 yrs δ_2	Overall δ	Power (%)		
				Unstratified	MN	**MR**
Scenario 1	0%	0%	0%	90	92	**94**
Scenario 2	-1%	-1%	-1%	83	85	**88**
Scenario 3	-1%	-2%	-1.5%	78	80	**85**
Scenario 4	-2%	-1%	-1.5%	78	80	**83**

100,000 simulations; all methods control the type 1 error rate (not shown)

The stratified **MR method** is most powerful in every scenario

Motivating NI Example: Results

$$H_{null}: \delta \le -10\% \quad \text{vs.} \quad H_{alt}: \delta > -10\%$$

δ = true overall treatment difference (A-B)

Observed percentage of responders				**Stratum Weights**	
Stratum	New Trt [A]	Old Trt [B]	A – B	MN method	MR method
< 65 yrs	91.1% (123/135)	91.1% (123/135)	0%	.40	.49
≥ 65 yrs	65.9% (135/205)	73.2% (150/205)	-7.3%	.60	.51

Method	Estimated δ	SD of estimated δ	95% CI for δ	Conclude Non-Inf?
MN	-4.4%	3.1%	(-10.5, 1.6)	No
MR	-3.8%	2.9%	(-9.4, 1.9)	Yes

Another NI Example

$$H_{null}: \delta \leq -10\% \quad \text{vs.} \quad H_{alt}: \delta > -10\%$$

Observed percentage of responders				**Stratum Weights**	
Clinical Syndrome	New Trt [A]	Old Trt [B]	A – B	MN method	MR method
Cellulitis	69.7% (134/192)	76.9% (147/191)	-7.2%	.64	.62
MCA	95.0% (38/40)	85.6% (33/39)	10.4%	.13	.15
Wound inf.	77.9% (53/68)	80.9% (55/68)	-2.9%	.23	.23

Method	Estimated δ	SD of estimated δ	95% CI for δ	Conclude Non-Inf?
MN	-3.9%	3.4%	(-10.6, 2.8)	No
MR	-3.6%	3.3%	(-10.2, 3.0)	No #

adding 1 responder to A in any stratum tips this to Yes!

Binary Endpoint

Treatment by Subgroup Interaction

Assessing consistency of treatment effect across subgroups

Lancet 2016; 387: 760–69

Adjusted treatment difference was calculated by a stratified Cochran-Mantel-Haenszel method with the randomisation strata of geographical region, allogeneic haemopoietic stem cell transplantation status, and active malignancy status. The 95% CI for the adjusted treatment difference was calculated on the basis of a normal approximation. Treatment-by-subgroup interaction (age, sex, race, ethnic origin, baseline neutropenic status, body-mass index, glomerular filtration rate, and enrolment period) was evaluated using a logistic regression according to the prespecified statistical significance value of $p<0 \cdot 15$. For assessment of

This is very common practice!

Assessing consistency of treatment effect across subgroups [2]

	Hypothetical Data		Difference
	Trt A	Trt B	(A-B)
Strat 1	98.3%	90.0%	8.3%
Strat 2	58.3%	50.0%	8.3%

	Hypothetical Data		Difference
	Trt A	Trt B	(A-B)
Strat 1	97.8%	94.4%	3.4%
Strat 2	70.0%	47.5%	22.5%

Mehrotra DV (2001; Drug Information Journal).

$$OR = \frac{p_A}{1-p_A} \Big/ \frac{p_B}{1-p_B}$$

Assessing consistency of treatment effect across subgroups [2]

Logistic regression is appropriate for assessing consistency across subgroups for **odds ratios** but NOT for **differences**!

NO treatment by stratum interaction on difference scale

	Hypothetical Data		Difference	Odds Ratio
	Trt A	Trt B	(A-B)	(A:B)
Strat 1	98.3%	90.0%	8.3%	6.5
Strat 2	58.3%	50.0%	8.3%	1.4

NO treatment by stratum interaction on odds ratio scale

	Hypothetical Data		Difference	Odds Ratio
	Trt A	Trt B	(A-B)	(A:B)
Strat 1	97.8%	94.4%	3.4%	2.6
Strat 2	70.0%	47.5%	22.5%	2.6

Mehrotra DV (2001; Drug Information Journal).

$$OR = \frac{p_A}{1-p_A} \Big/ \frac{p_B}{1-p_B}$$

Assessing Treatment by Subgroup Interaction

Metric: difference in treatment response rates

- p_{ij} = true response rate for subgrp *i*, trt *j* $(i = 1,2,...s;\ j = A,B)$
- $\delta_i = p_{iA} - p_{iB}$ (true treatment difference in subgrp *i*)
- $H_0 : \delta_1 = \delta_2 = \ldots = \delta_s$ (no interaction on difference scale)
- Methods
  - Linear probability model (GENMOD: dist=binomial, link=identity)
  - Fleiss (1981), Gail & Simon (1985), DerSimonian & Laird (1986)*
  - Mehrotra (1997, 2001)*

  * plug-in methods using observed proportions

Time-to-Event Endpoint

Stratified Time-to-Event Analysis: Methods

Examples of "events": death, cancer progression, virologic failure, etc.

- **Traditional Method (1-step method)**
  - Stratified Cox PH model (or stratified logrank test)
  - Assumes the true hazard ratios (HRs) are the same for all strata; this is rarely true!

- New Method (2-step method)*
  - Fit Cox PH model within each stratum $i \Rightarrow \hat{\beta}_i, V(\hat{\beta}_i)$
  - Combine stratum-specific estimates using MR or SSIZE wts

    Stratified Cox model implicitly uses INVAR wts $\{V(\hat{\beta}_i)\}^{-1}$
  - Does NOT assume a constant HR across strata

* Mehrotra, Su, Li (2012, Statistics in Medicine)

Merck Vaccine Trial (2012; design stage)

- Placebo vs. V212 (shingles vaccine) in chemotherapy patients
- Total N ≈ 6000, 2 strata (unequal baseline risk), 1:1 (P:V)
- **Success**: 95% CI for **overall** vaccine efficacy (VE) is > **25%**

"super-superiority"

Simulated True VEs		**Statistical Power (%)**	
Stratum 1	Stratum 2	Stratified Cox model	New Method (2-step; MR wts)
VE_1=50%	VE_2=50%	90	90
VE_1=60%	VE_2=38%	76	85
VE_1=63%	VE_2=32%	69	85

5,000 simulations; equal stratum sizes; $VE = 1-[h_V(t)/h_P(t)]$

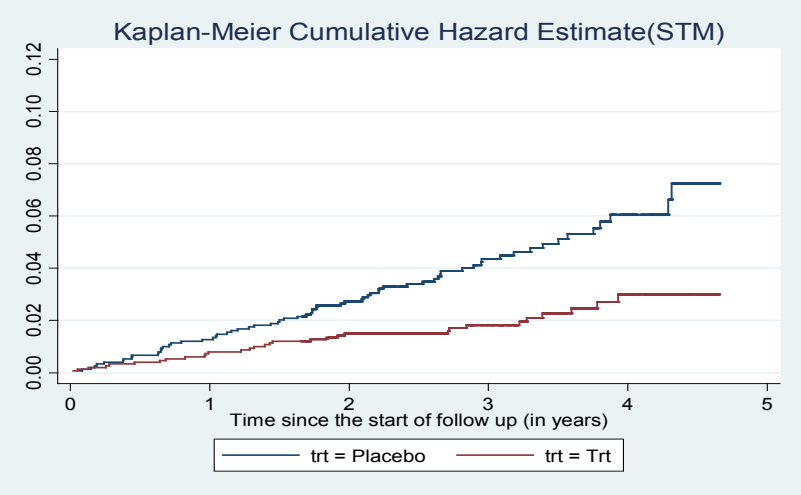
- Use of the traditional method (stratified Cox model) would require > 30% increase in N to ensure ≥ 85% power

Merck Vaccine Trial (continued)

Hypothetical Results

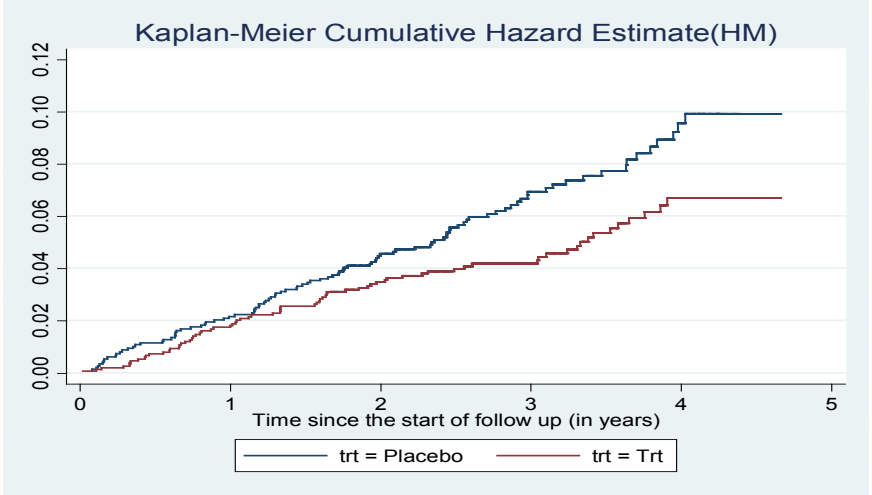
Stratum 1

VE_{obs}=54.6%



Stratum 2

VE_{obs}=31.0%



Method of Analysis	Overall VE Estimate	95% CI	Lower Bound
Stratified Cox PH model	39.7%	(23.0%, 52.7%)	LB < 25%
New method (2-step; MR wts)	42.4%	(25.8%, 55.3%)	LB > 25%

MR weights: 0.55 (stratum 1) and 0.45 (stratum 2)

Conclusions: Topic #5

- To improve the POS for a randomized clinical trial:
  - Stratify randomization by **key prognostic factor(s)**
  - Use **simulations** to guide selection of the primary method for assessing the **overall** treatment effect
  - TTE endpoint: **2-step analysis** is a robust alternative to the stratified Cox model analysis
- When assessing consistency of treatment effect across strata, test for **interaction** on the appropriate **scale**
  - Plug-in methods can be used for binary and TTE endpoints; they have good properties and patient-level data are not required

References: Topic #5

- Farrington CP, Manning G (1990). Statistics in Medicine, 9, 1447 1454
- Mantel N, Haenszel W (1959). Journal of the Natational Cancer Institute, 22, 719-748
- Mehrotra DV, Railkar R (2000). Statistics in Medicine, 19, 811-825
- Mehrotra DV (2001). Drug Information Journal, 35, 1343-1350
- Mehrotra DV, Su S, Li X (2012). Statistics in Medicine, 31, 1849-1856
- Miettinen O, Nurminen M (1985). Statistics in Medicine, 4, 213-226.

Topic #6

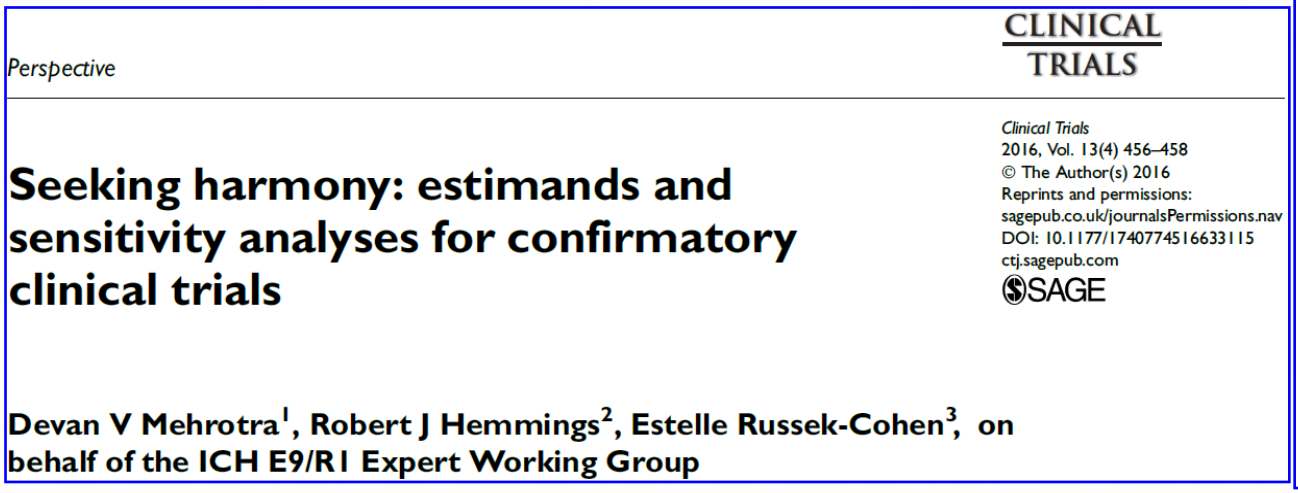
Estimand-Aligned Primary and Sensitivity Analyses
ICH E9/R1

ICH E9/R1: Background

- 1998: ICH E9 (Statistical Principles for Clinical Trials)

  Foundational: randomization, double blind, interim analysis, non-inferiority, etc.

- 2014: regulatory statisticians proposed creation of an expert working group (EWG) to develop an E9 addendum (E9/R1) on **estimands** and **sensitivity analyses**



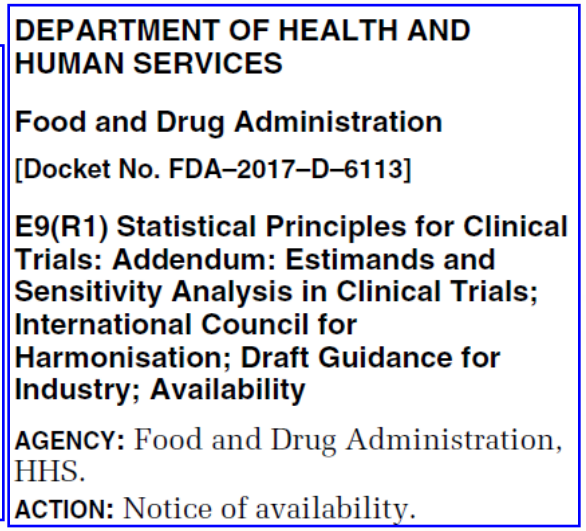
Perspective

CLINICAL TRIALS

Clinical Trials
2016, Vol. 13(4) 456–458
© The Author(s) 2016
Reprints and permissions:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/1740774516633115
ctj.sagepub.com
SAGE

Seeking harmony: estimands and sensitivity analyses for confirmatory clinical trials

Devan V Mehrotra[1], Robert J Hemmings[2], Estelle Russek-Cohen[3], on behalf of the ICH E9/R1 Expert Working Group



DEPARTMENT OF HEALTH AND HUMAN SERVICES

Food and Drug Administration

[Docket No. FDA–2017–D–6113]

E9(R1) Statistical Principles for Clinical Trials: Addendum: Estimands and Sensitivity Analysis in Clinical Trials; International Council for Harmonisation; Draft Guidance for Industry; Availability

AGENCY: Food and Drug Administration, HHS.

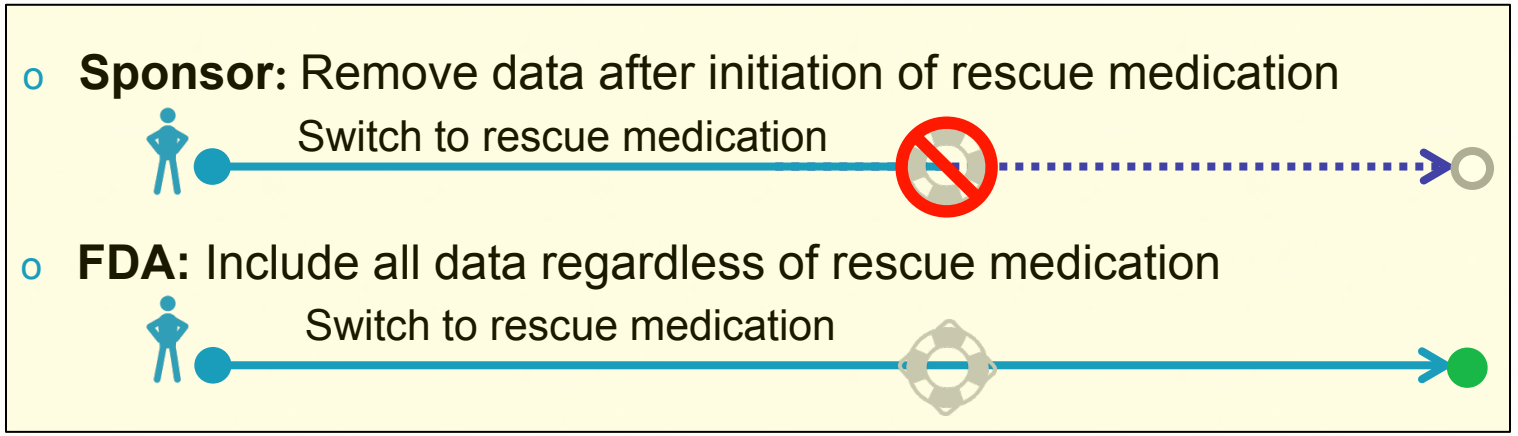
ACTION: Notice of availability.

- 2017: draft ICH E9/R1 released for public comment (> 2000 comments received)

ICH E9/R1: Why?

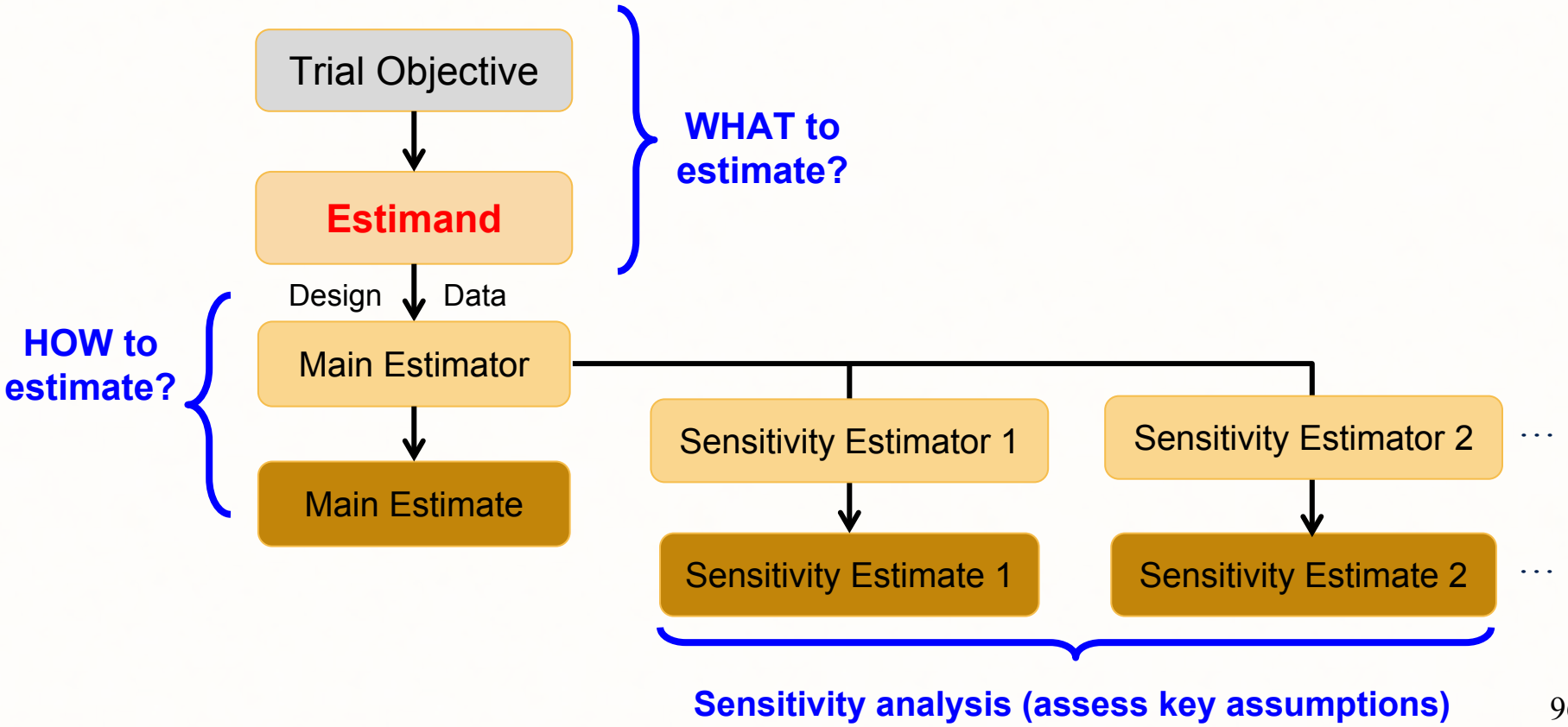
- **Feedback from regulatory statisticians** on several clinical trial protocols & new drug applications:
  - Insufficient clarity in objectives and related treatment effect parameters (i.e., estimands) of interest
  - Lack of logical connectivity between trial objectives, design, conduct, analysis and interpretation
  - Misalignment between "missing data" analysis methods and estimands of interest

2011 FDA advisory committee for dapagliflozin



ICH E9/R1: A Structured Framework

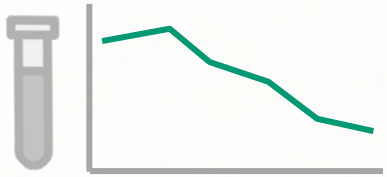
For a given trial objective: aligning target of estimation, design, method of estimation and sensitivity analysis



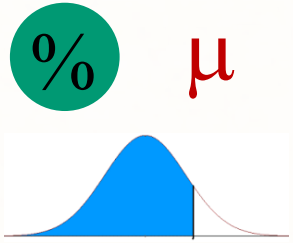
ICH E9/R1: Inputs for Defining an Estimand

A
Population

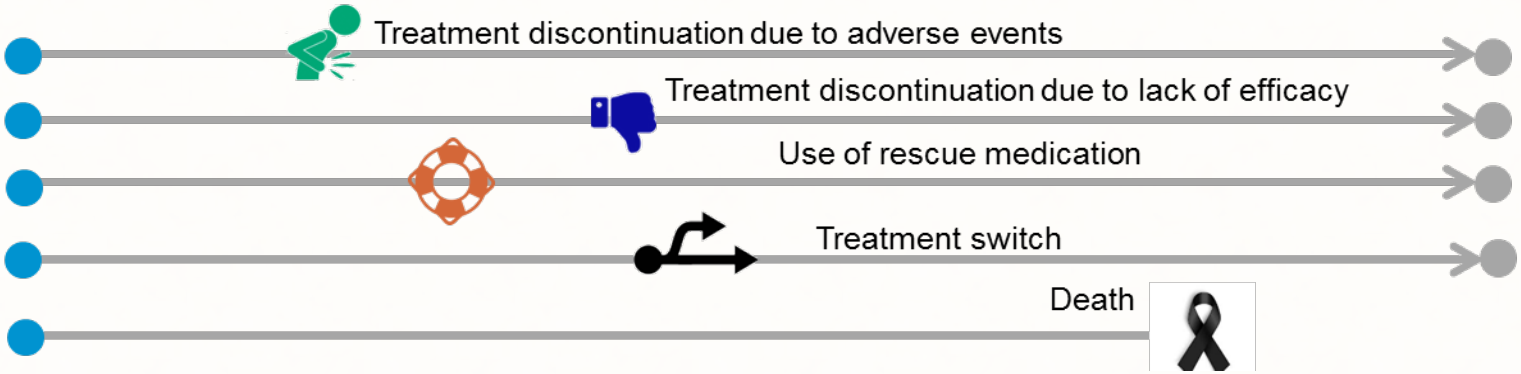
B
Variable (Endpoint)



D
Population-level
Endpoint Summary



C
Intercurrent Event(s)



Case Study: Diabetes (HbA1c; asymptomatic)

Primary Objective: assess whether drug is more effective than placebo in lowering HbA1c without rescue medication (the latter was allowed, but to address a different objective)

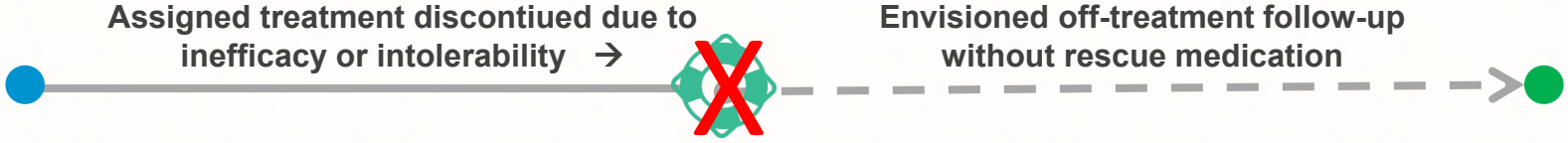
Construction of Primary Estimand

Population: adults with type II diabetes (per intended label)

Endpoint: HbA1c change from baseline at 24 weeks

Intercurrent event: rescue medication

Treatment effect of interest is based on envisioned endpoint under complete follow-up even if the assigned treatment is discontinued, but without rescue medication at any time



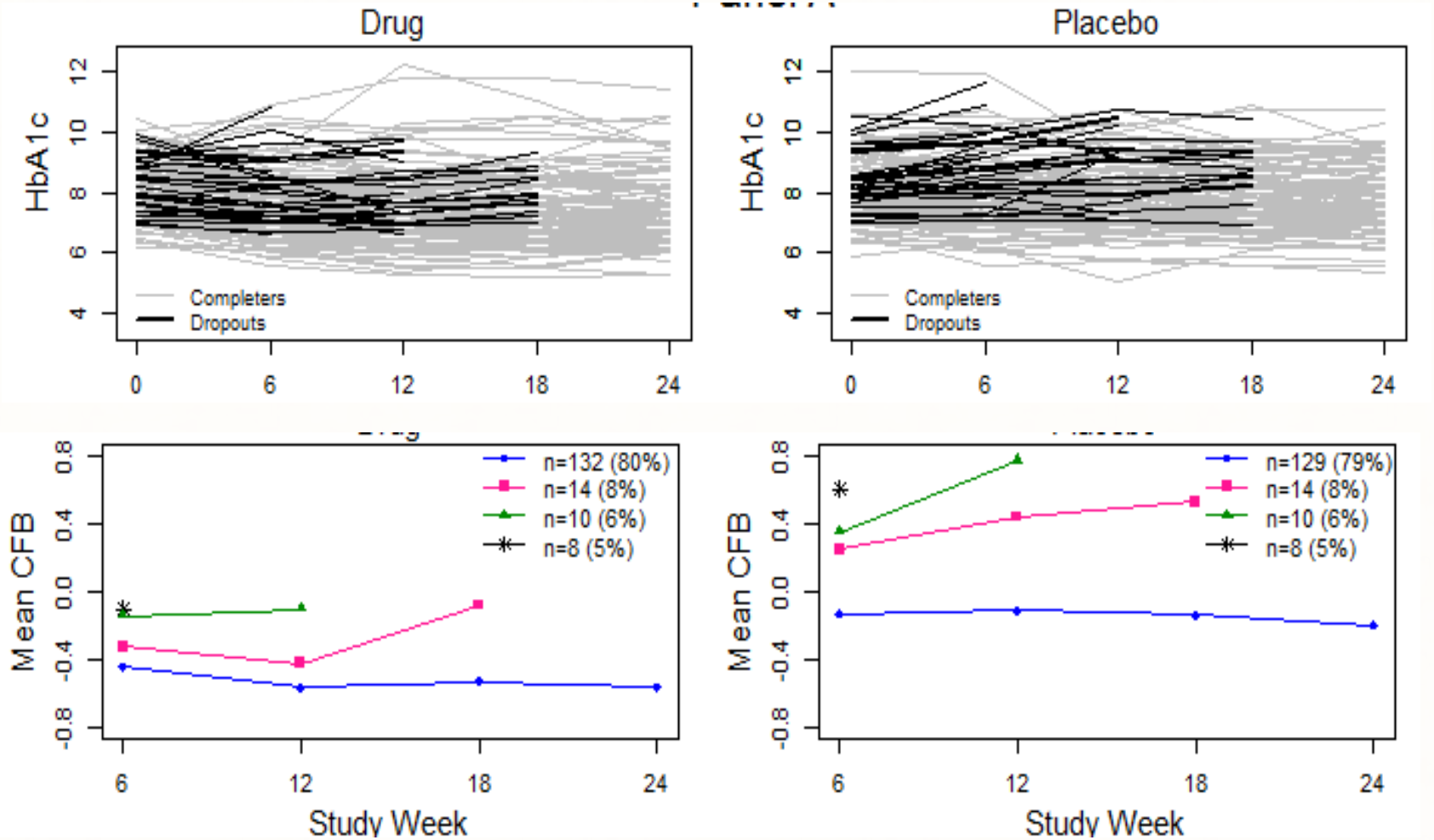
Population-level summary: mean of the endpoint

Case Study (continued)

- **Estimand**: between-treatment difference in target population endpoint means for the treatment effect of interest (δ)
- **Statistical objectives**
  - Deliver acceptable point estimate and 95% CI for δ
  - Test H_{null}: $\delta=0$ vs. H_{alt}: $\delta<0$ (with type 1 error rate $\leq \alpha$)
- **Tackling rescue medication in the analysis**
  - Given estimand of interest, HbA1c values after initiation of rescue medication can be discarded, resulting in "missing" endpoint data for such patients (and dropouts)

- **Analysis challenge**: all randomized patients need to be included in the analysis (per the estimand), so how do we tackle the missing endpoint data problem?

Case Study (continued)



% missing endpoint: **20%** (33/165) Drug **21%** (34/164) Placebo

1 patient assigned to drug and 2 patients assigned to placebo were dropouts before week 6

Case Study (continued)

Control-based mean imputation is one way to tackle the missing data problem (to be pre-specified in the analysis plan)

- Obs = endpoint observed, miss = endpoint missing
- π_i^{miss} = true Pr(endpoint missing under trt i) = $1 - \pi_i^{obs}$

Placebo	***Drug***
$\mu_P = \pi_P^{obs}\,\mu_P^{obs} + \pi_P^{miss}\,\mu_P^{miss}$	$\mu_D = \pi_D^{obs}\,\mu_D^{obs} + \pi_D^{miss}\,\mu_D^{miss}$
$\hat{\mu}_P = \hat{\pi}_P^{obs}\,\hat{\mu}_P^{obs} + \hat{\pi}_P^{miss}\,\hat{\mu}_P^{miss}$	$\hat{\mu}_D[c] = \hat{\pi}_D^{obs}\,\hat{\mu}_D^{obs} + \hat{\pi}_D^{miss}\,(\boldsymbol{\hat{\mu}_P + c})$

$\hat{\mu}_P^{miss}$ = estimate of μ_P assuming missing endpoints are MAR for placebo

Estimand: $\delta = \mu_D - \mu_P$

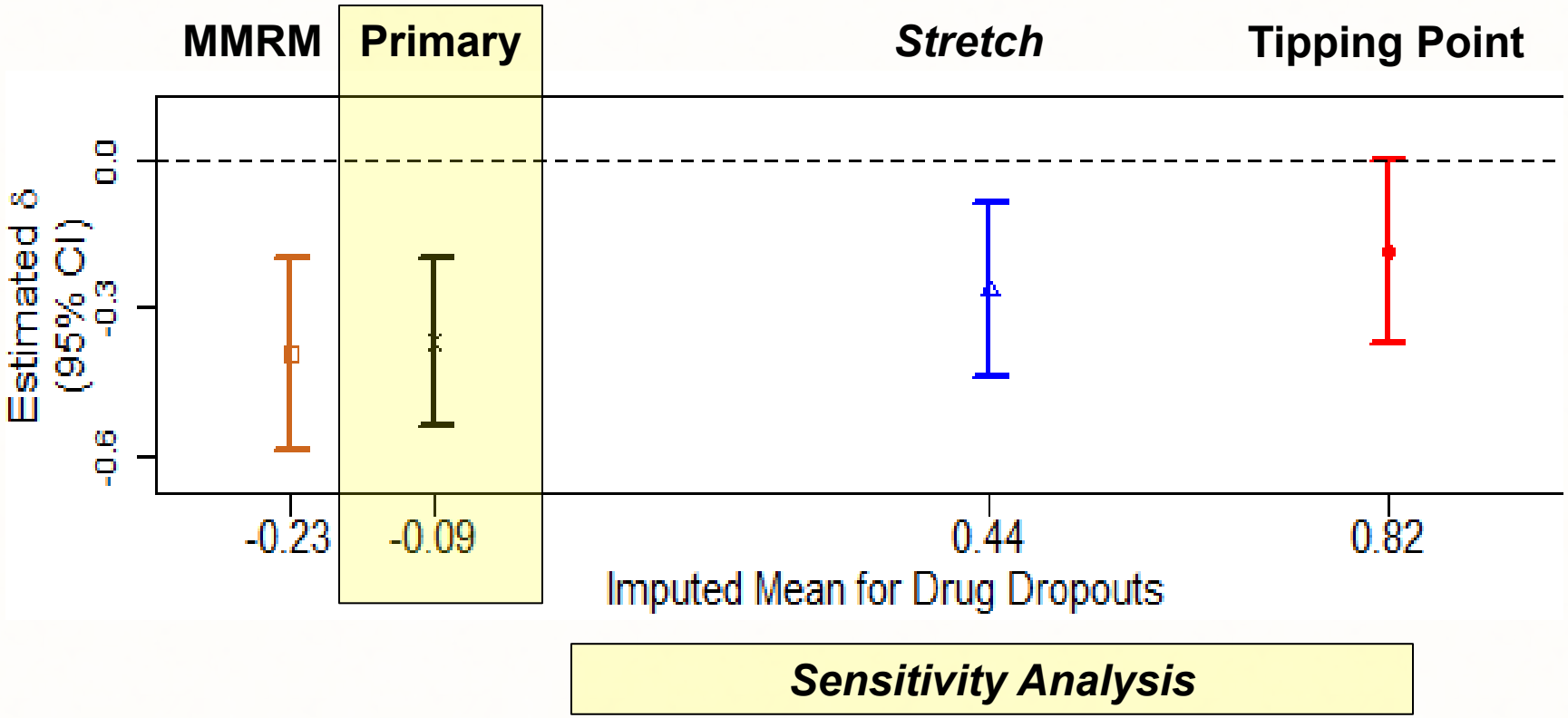
Estimation: $\hat{\delta}[c] = \hat{\mu}_D[c] - \hat{\mu}_P$

Primary Analysis: use $c = 0$

***Sensitivity Analysis*:** increase c until **Tipping Point** is reached

Details: Mehrotra, Liu, Permutt (2017; Pharmaceutical Statistics)

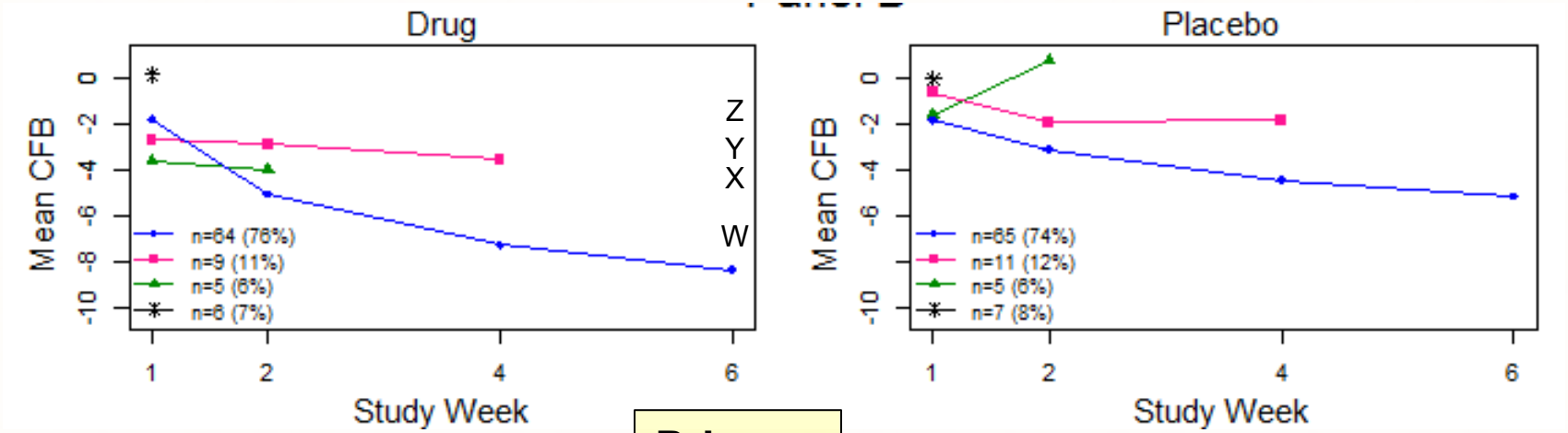
Case Study (continued)

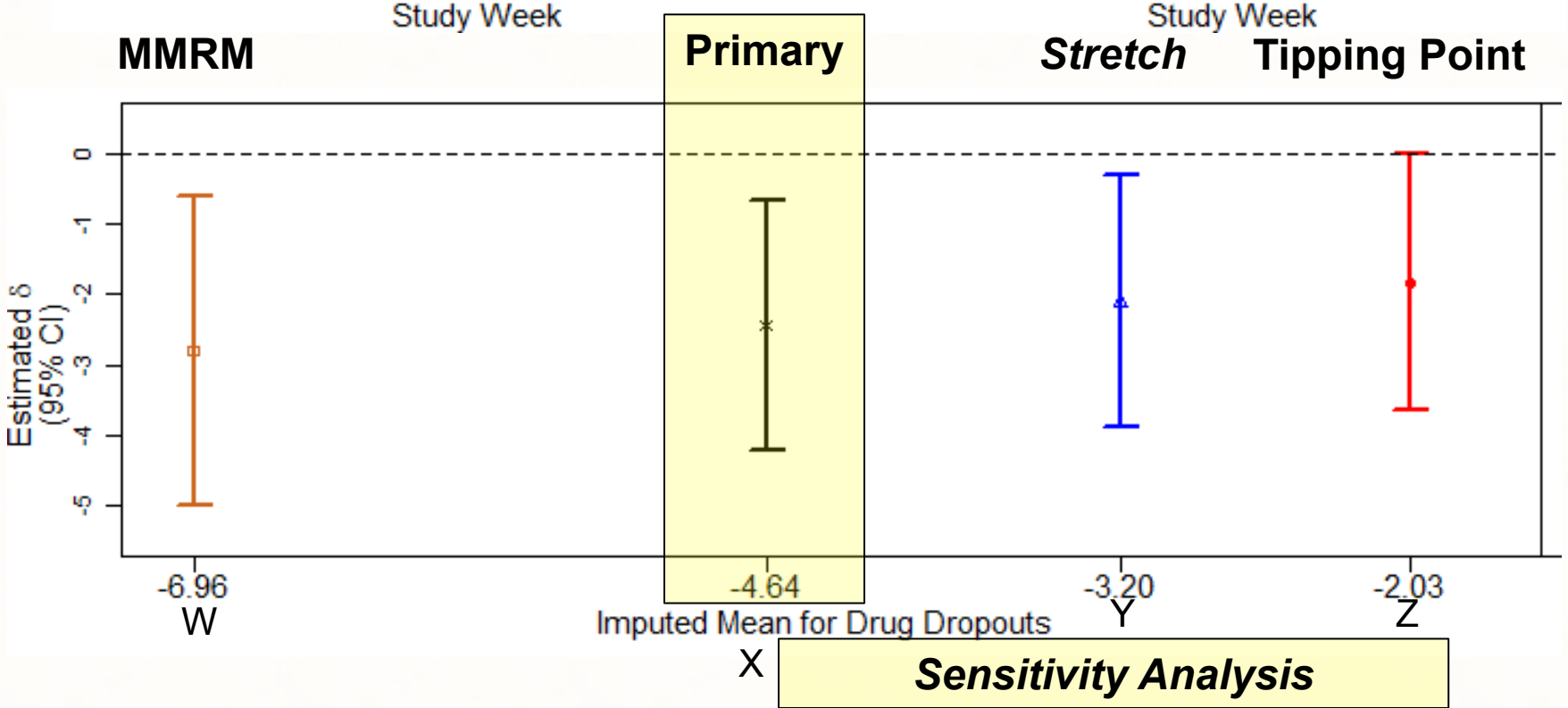


- MMRM: mixed model repeated measures assuming MAR dropout for drug and placebo [shown for historical reference only]
- *Stretch*: imputed mean for drug dropouts matches estimated mean for placebo dropouts assuming MAR dropout for placebo
- Tipping point after *stretch* imputation ⇒ **robust** evidence of drug effect

Case Study #2: Depression (HAMD17; symptomatic)

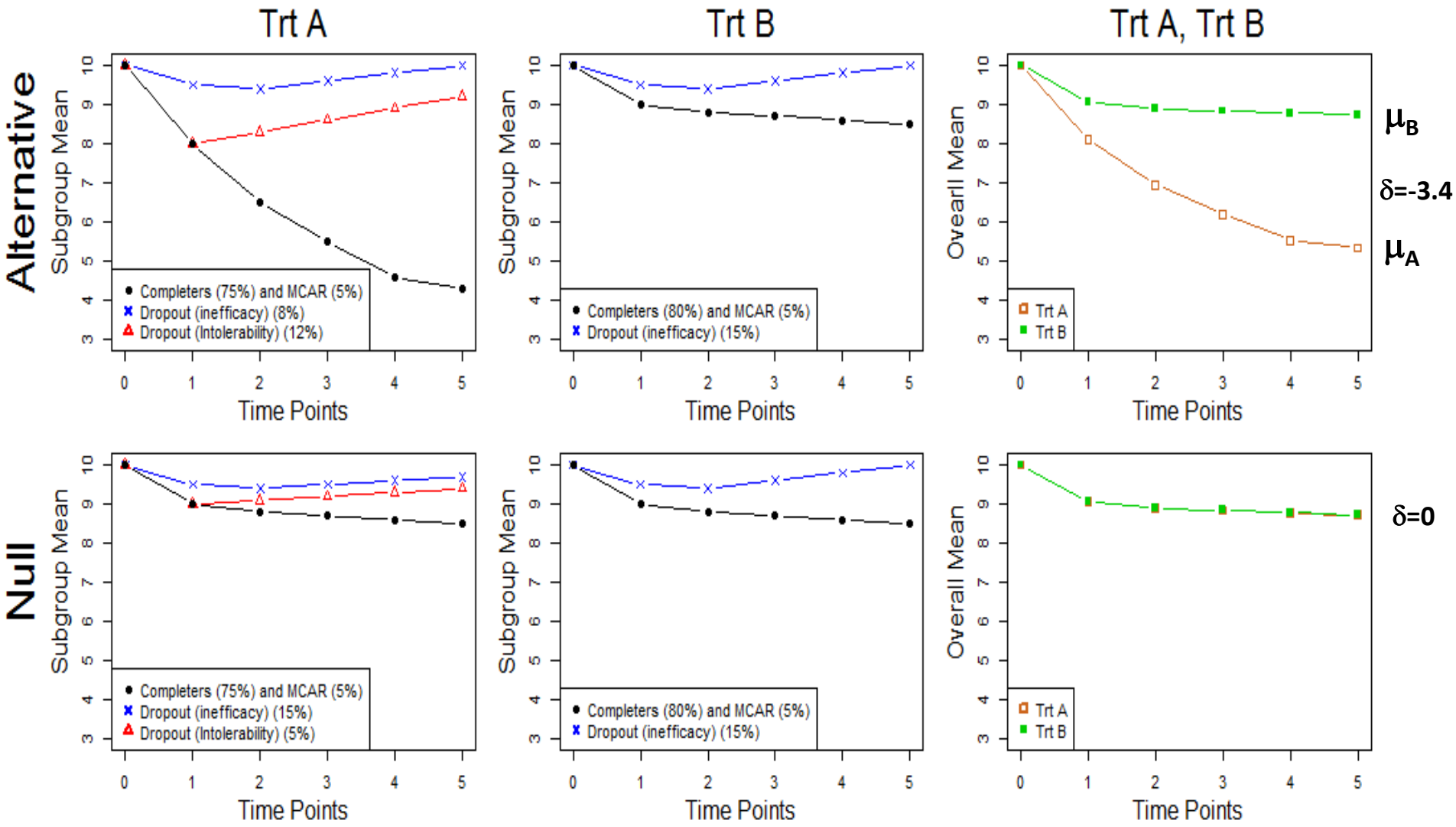
% missing endpoint: **24%** (20/84) Drug, **26%** (23/88) Placebo





Simulation Setup: True Longitudinal Mean Profiles

N = 100/group, AR1(ρ=0.85) correlation structure within each pattern



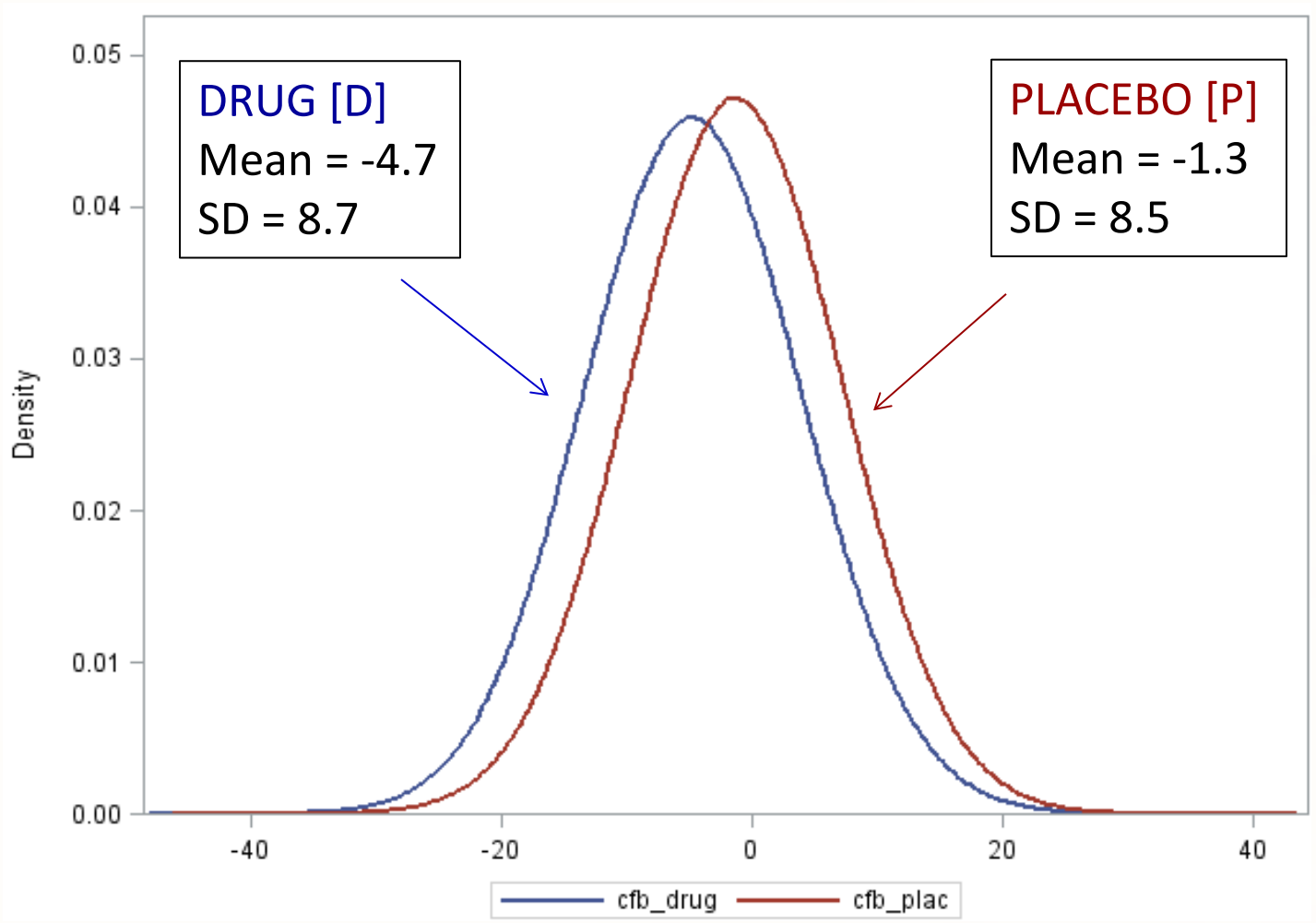
Simulated study dropouts[A=drug, P=placebo]

Intolerability: at time point 1; Inefficacy: at time point 2

Dropout
A: 25%
B: 20%

Alternative Hypothesis: True Endpoint Distributions for Drug, Placebo

Endpoint = true change from baseline at time point 5



Note: each is a mixture of normal distributions (not visually obvious!)

Estimand: δ = -3.4 (true difference, D-P)

Simulation Results: Under Null Hypothesis

Method	Bias	Type 1 Error (%) Favoring Drug (Target ≤ 2.5%)	95% CI Error Favoring Drug (Target ≤ 2.5%)
MMRM	-0.01	2.6	2.6
Proposed Primary Analysis	-0.05	2.5	2.5
Proposed Stretch Analysis	0.01	2.2	2.2

Simulation Results: Under Alternative Hypothesis

Method	% Bias	Power (%)	95% CI Error Favoring Drug (Target ≤ 2.5%)
MMRM	-21.5	~~95.2~~	9.4
Proposed Primary Analysis	5.4	95.3	1.2
Proposed Stretch Analysis	7.2	94.1	1.1

Negative bias means drug benefit (vs. placebo) is overestimated; 5,000 simulations

HR Estimands for Time-to-Event Endpoints

Estimand for a homogeneous population

- $S_j(t) =$ true proportion event-free at time t under trt j

  Under proportional hazards: $\theta(t) = \frac{\log\{S_A(t)\}}{\log\{S_B(t)\}} = \theta$ for all t

  θ is the time-invariant hazard ratio under PH

Estimand for a mixture of homogeneous subpopulations

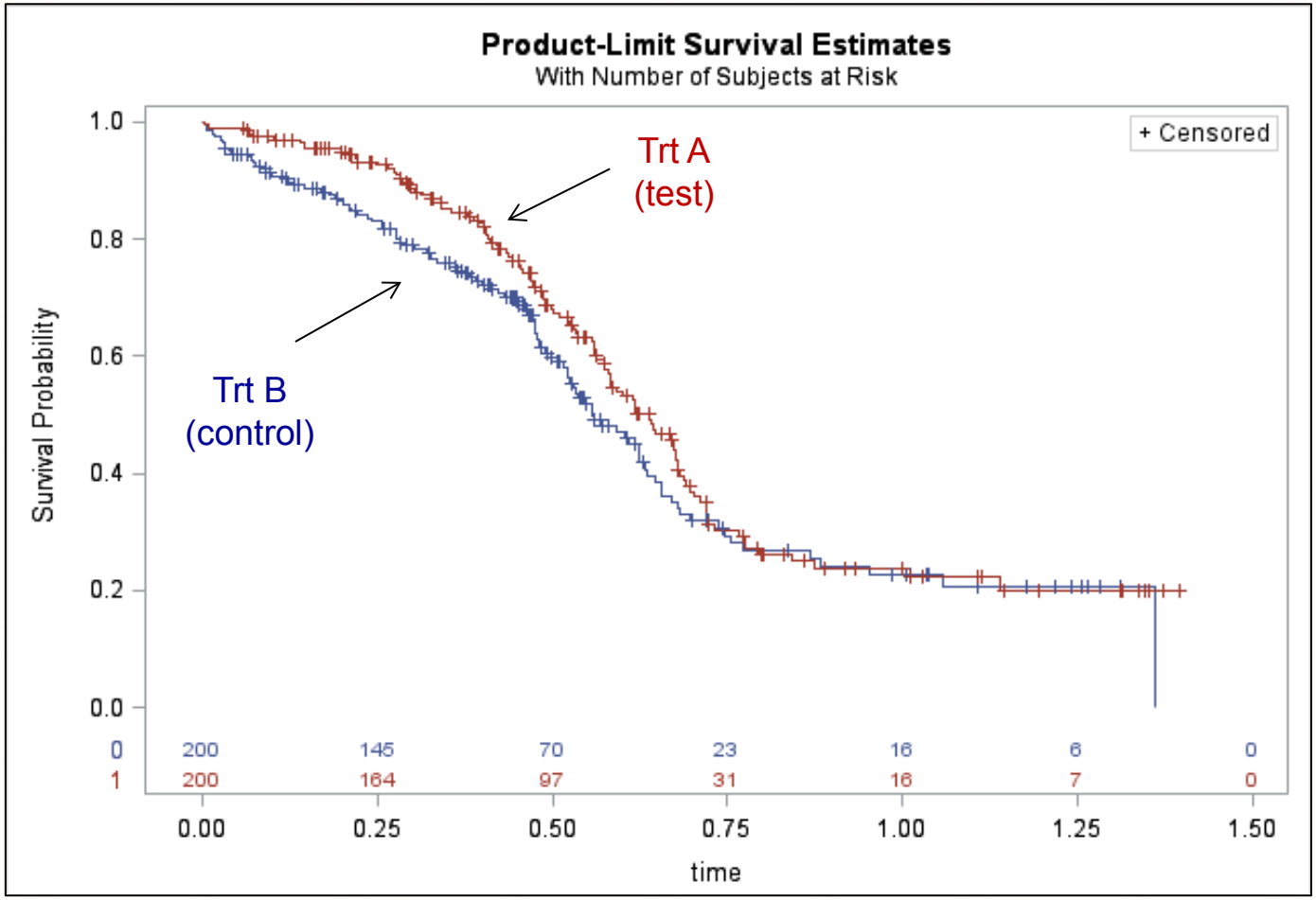
- Assume PH holds within each subpopulation ("stratum")

  Define $\bar{\beta} = \sum_j f_j \beta_j$

  f_j = prevalence and $\beta_j = \log(\theta_j)$ for stratum j

  $\bar{\theta} = \exp(\bar{\beta})$ is the time-invariant average hazard ratio

Example: Simulated Dataset (N=400)

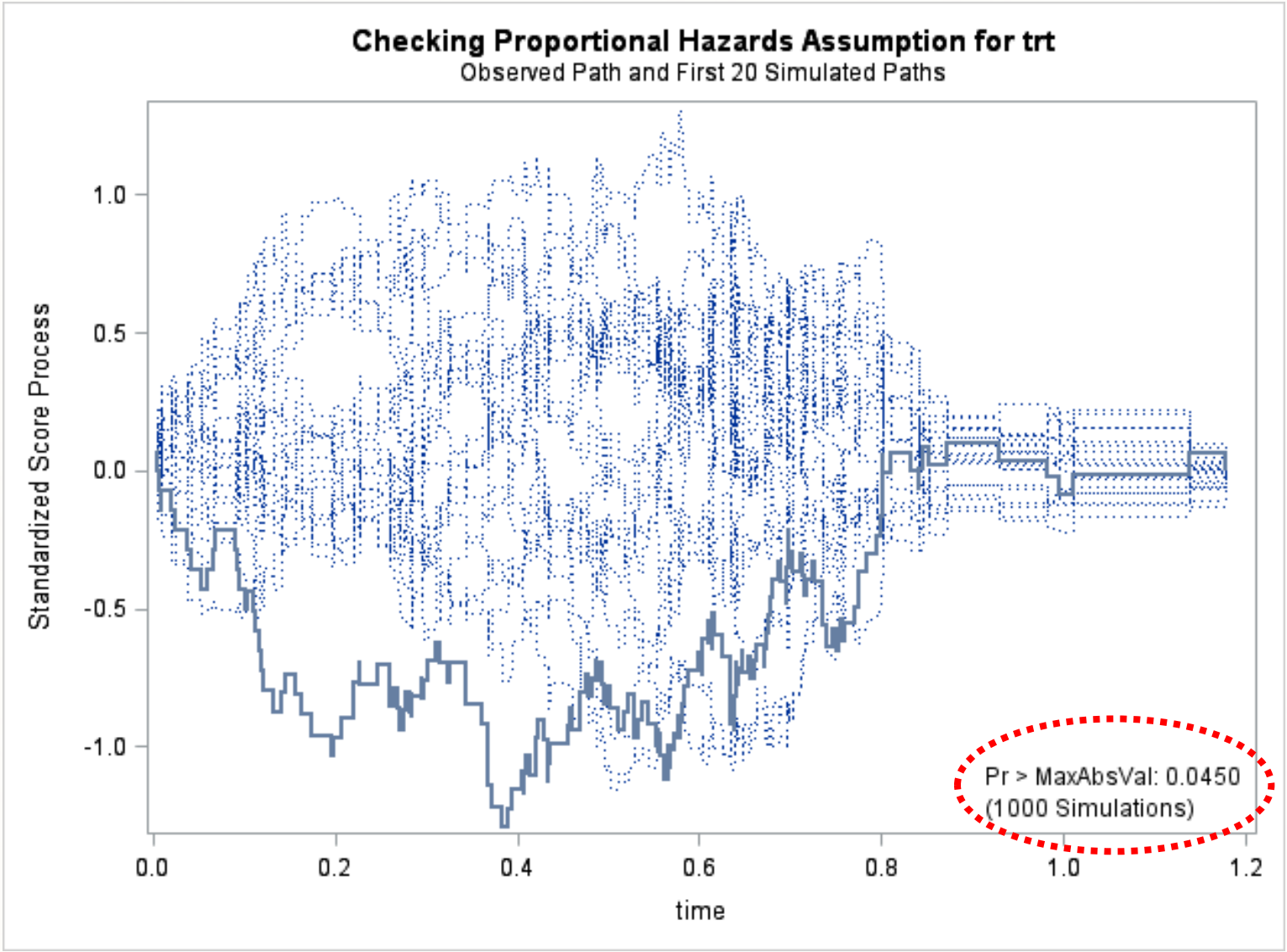


Method	Hazard Ratio (HR)		2-tailed p-value
	Estimate	95% CI	
Cox PH model	0.82	(0.62, 1.07)	.1398

Are results reliable?

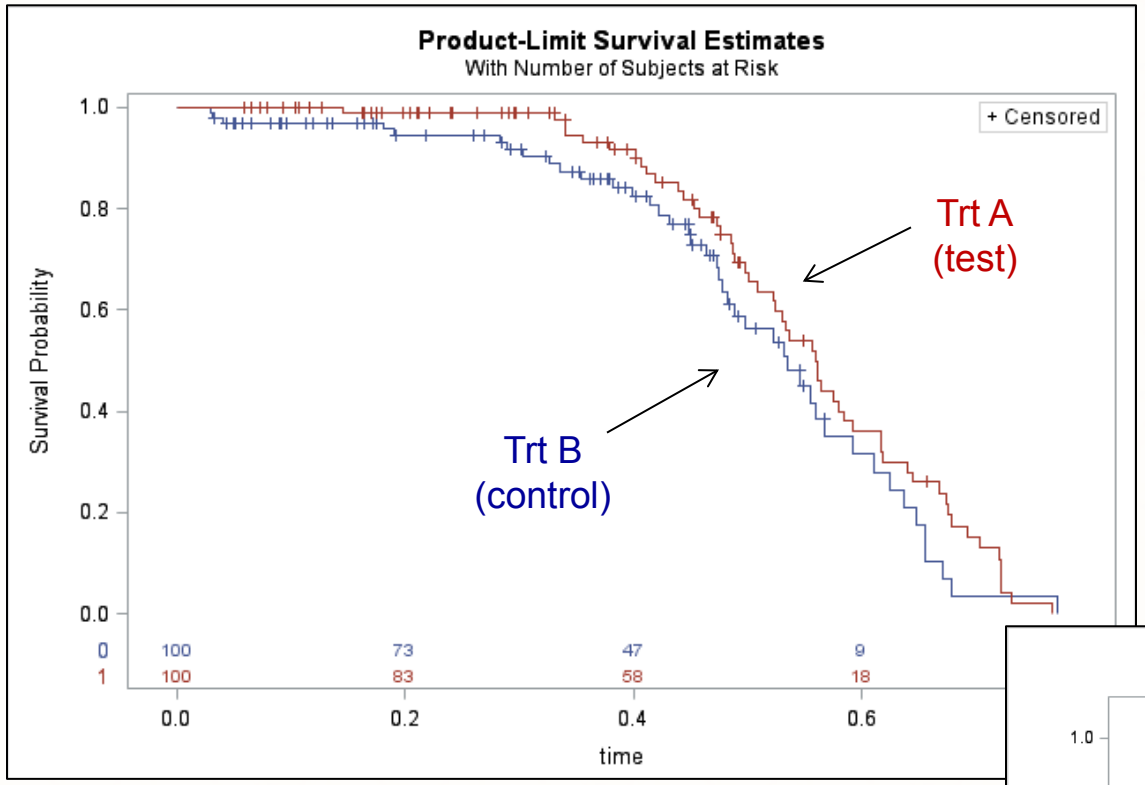
Example (continued): Check PH assumption

Using Cumulative Sums of Martingale Residuals (Lin et al 1993; PHREG/ASSESS)



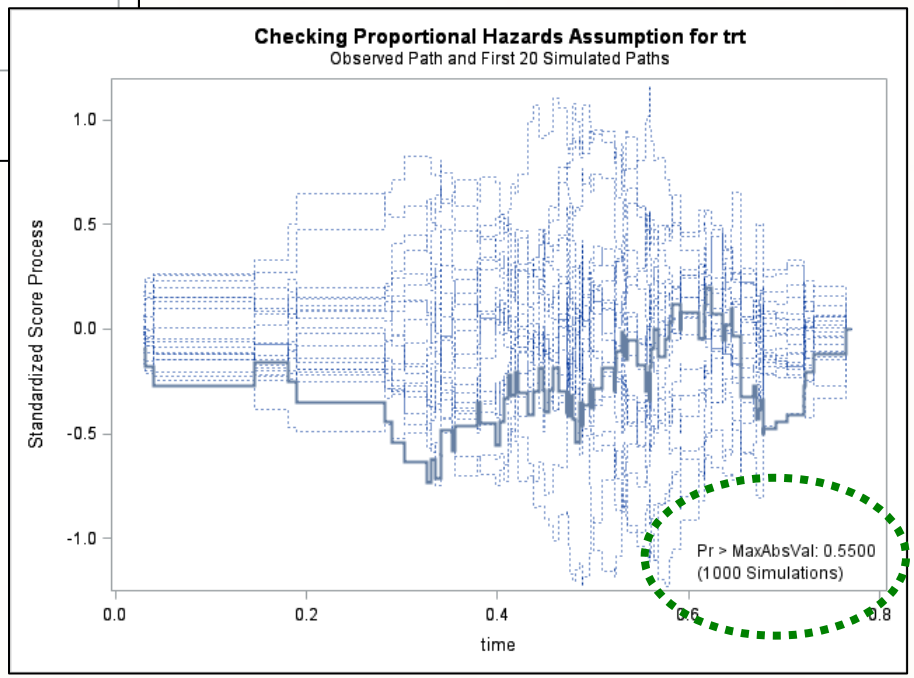
Not PH

Example (continued): Patients in Baseline Risk "Stratum 1"

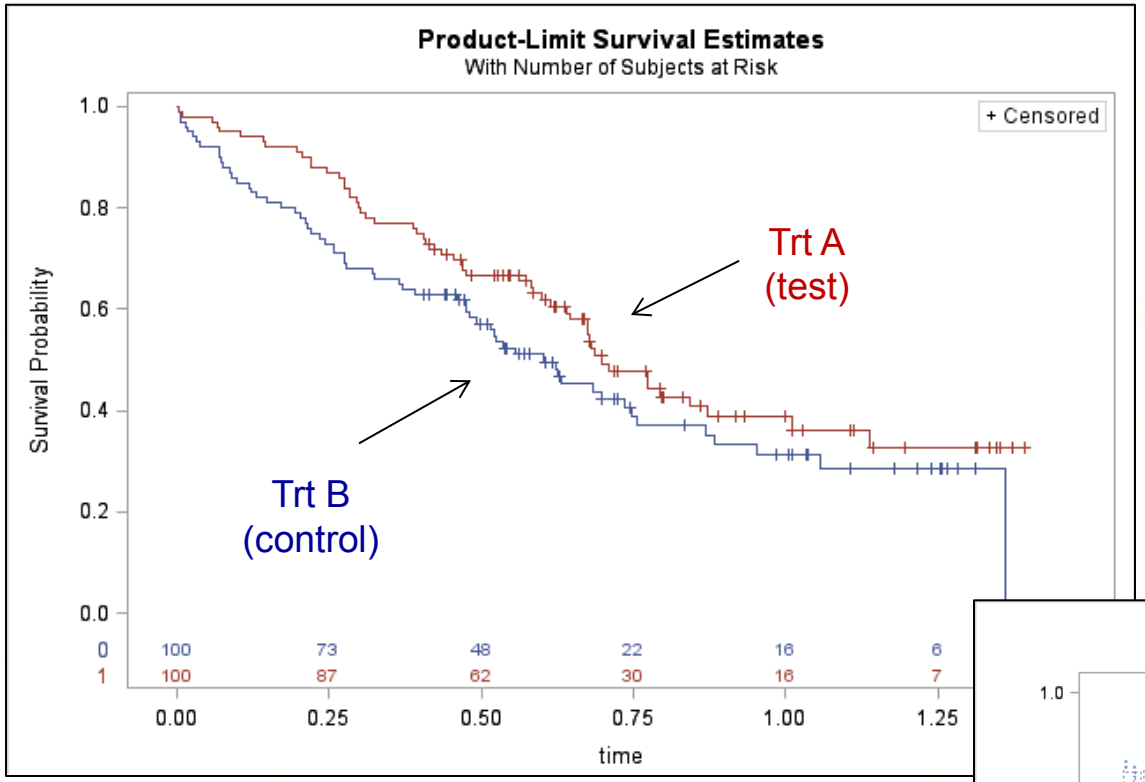


PH

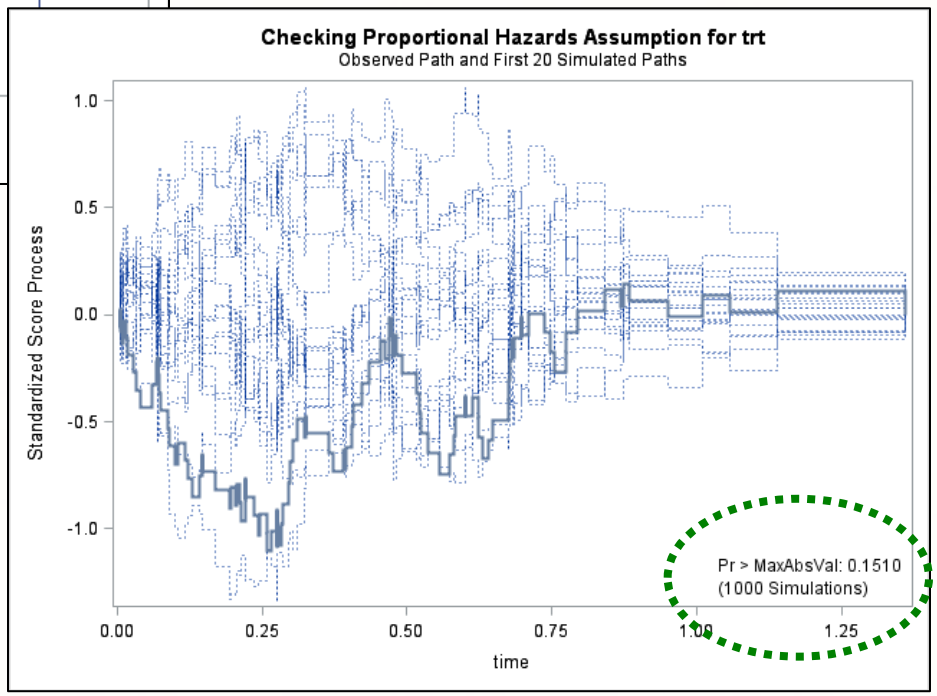
No. of patients	**HR in Stratum 1** Estimate	95% CI
200	0.74	(0.49, 1.13)



Example (continued): Patients in Baseline Risk "Stratum 2"



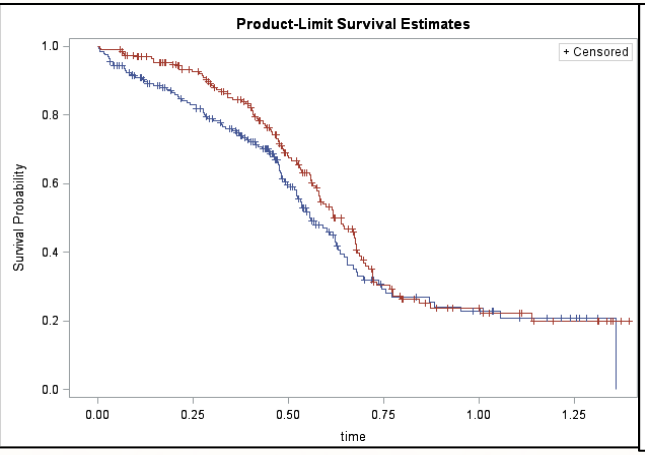
PH

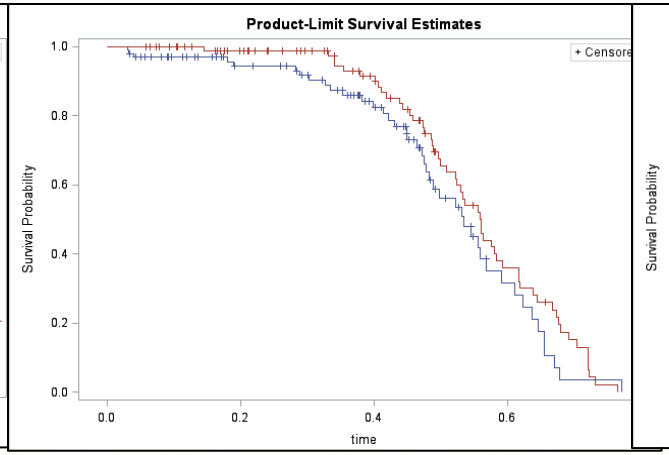


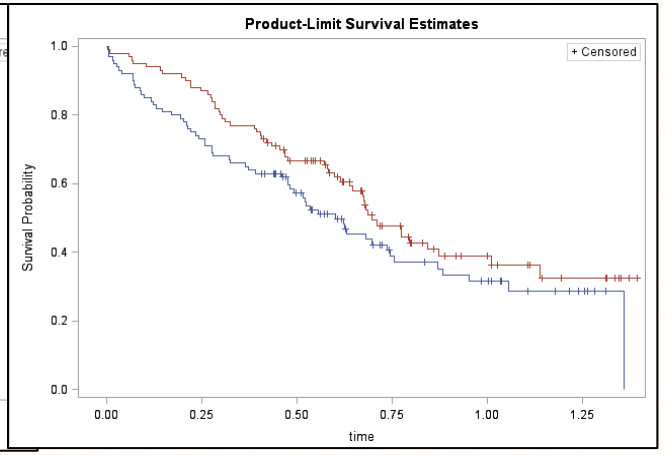
No. of patients	**HR in Stratum 2** Estimate	95% CI
200	0.74	(0.52, 1.07)

Example (continued): Summary of Results

Pooled (not PH) | Stratum 1 (PH) | Stratum 2 (PH)







HR=0.82 (interpretation?) | HR=0.74 | HR=0.74

Analysis Method	Hazard Ratio (HR) Estimate	95% CI	2-tailed p-value
Unstratified Cox PH (pooled)	0.82	(0.62, 1.07)	.1398
Stratified analysis (SSIZE wts)*	0.74	(0.56, 0.98)	.0362

* Mehrotra et al (2012; Statistics in Medicine)

Conclusions: Topic #6

- **ICH E9/R1 (estimands & sensitivity analyses) is intended to:**
  - Enable better planning of clinical trials and application dossiers for new drugs/vaccines/biologics
  - Strengthen understanding of decision-making by regulatory authorities and advisory committees

- **Open items for estimands and sensitivity analyses:**
  - Time-to-event endpoints under PH/non-PH settings
  - Non-inferiority trials: different from superiority?
  - Treatment effects in well-defined but non-identifiable populations (principal stratification)
  - Others

References: Topic #6

- Carpenter J, Roger J, Kenward M (2013). Journal of Biopharmaceutical Statistics, 23, 1352-71.
- LaVange L, Permutt T (2016). Statistics in Medicine, 35, 2853-2864.
- Mehrotra DV, Hemmings RJ, Russek-Cohen E (2016). Clinical Trials, 13, 456-458.
- Mehrotra DV, Liu F, Permutt T (2017). Pharmaceutical Statistics, 16, 378-392.
- Permutt T, Li F (2017). Pharmaceutical Statistics, 16, 20-28.

Course Wrap-Up

- We have discussed several examples of notably more powerful clinical trial analyses that enable improved operational efficiency across all phases of clinical development relative to traditional approaches

- Use of these newer methods can (subsequently) help deliver downstream benefits for patients, prescribers, payers, pricing and product labels

- SAS or R code for implementation is available from the instructor: devan_mehrotra@merck.com