Making Clinical Trial Results Robust

CAPS Methods Core Seminar
UCSF
March 11, 2016

Jitendra Ganju
Global Blood Therapeutics
jganju@globalbloodtx.com

Collaborators
Julie Ma (Gilead Sciences), Xinxin Yu (U. of Wisconsin)
and
Yunzhi Lin (Takeda Pharma), Kefei Zhou (Amgen)

Clinical trial results are not robust because we've built fragility into it

Why are Trial Results not Robust?

- **How do you invest your savings?**
  - Invested in shares of one company or do you diversify?
- Why do you hedge your bets?
- With analysis of clinical trial data no formal plan to minimize risk

Pre-specification

- We pre-specify everything

  Detailed description of primary and secondary endpoints (EPs), rules for handling missing data, multiple comparisons and control of overall error rate, method of analysis for each endpoint,…

- Usually one primary endpoint

- **Always just one statistic to formally analyze endpoints**

  Like investing your savings in shares of one company

  What if you got it wrong? Three examples next

Absolute change, relative change; Data transformations; Effects of covariates on hazard function multiplicative or additive; Covariate adjustment in non-linear models

A Randomized Trial of Propranolol in Patients With Acute Myocardial Infarction

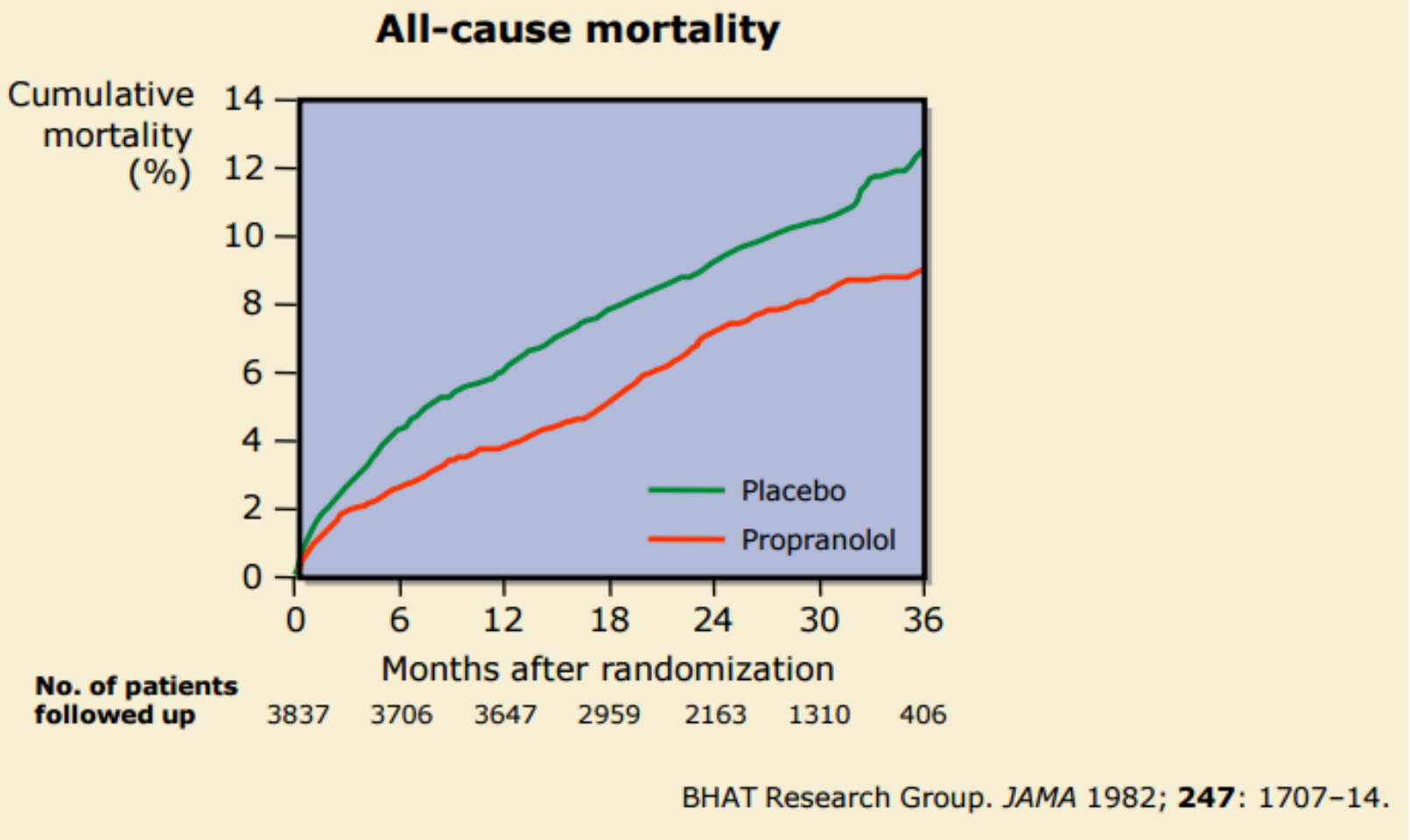
I. Mortality Results

β-Blocker Heart Attack Trial Research Group

BHAT 1978-1981
JAMA 1982

- Randomized, double-blind, propranolol vs placebo
  N = ~3800, 30-69 yrs, hospitalized with prior acute MI
- Primary endpoint: all-cause mortality
  Test statistic: logrank
- Study halted early. Mortality less in treatment group

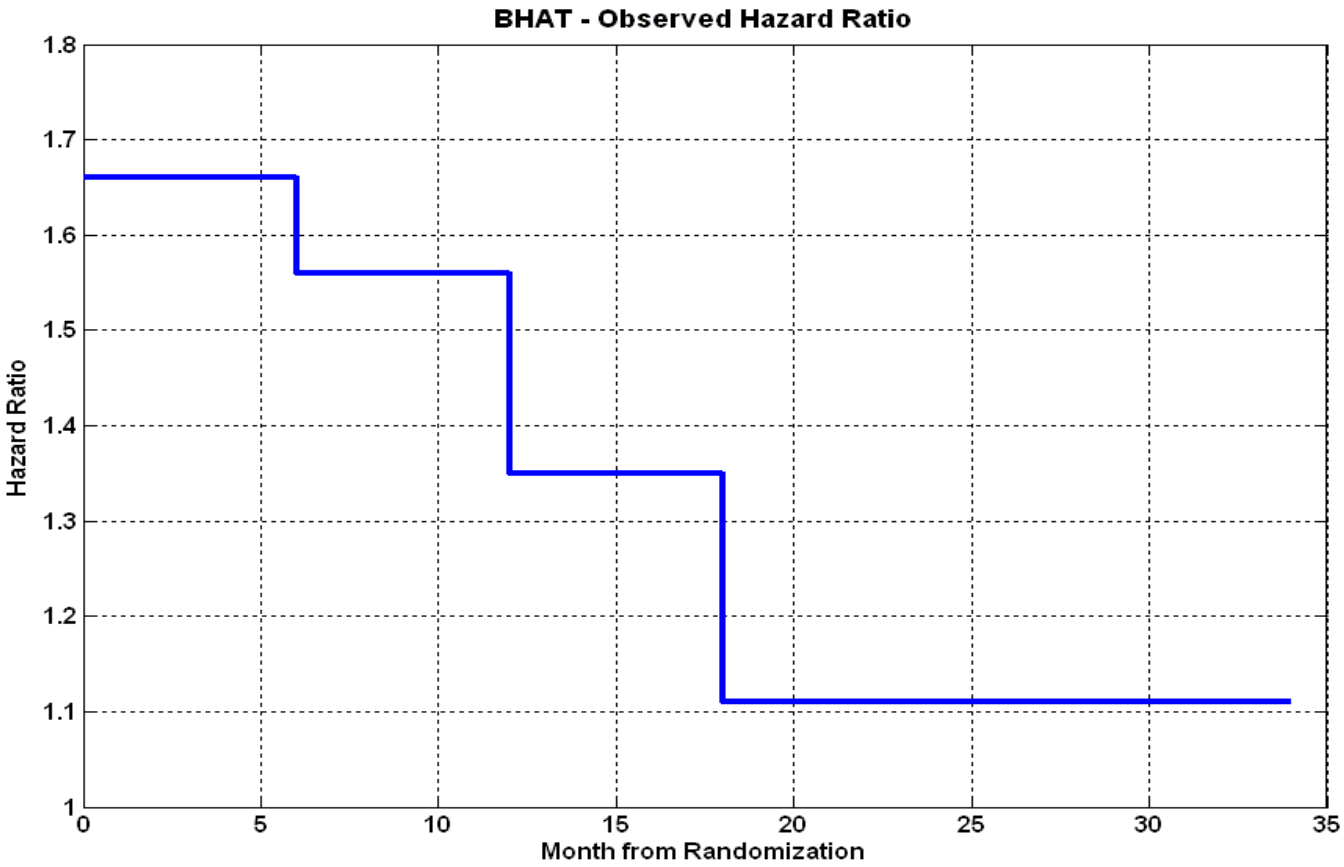
BHAT



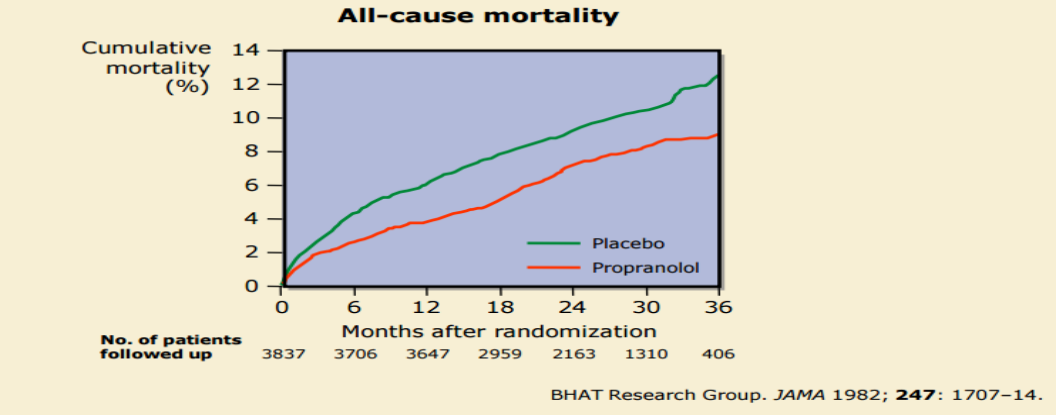
http://solat.cl/imgsolat/archivoarticulos/2.pdf

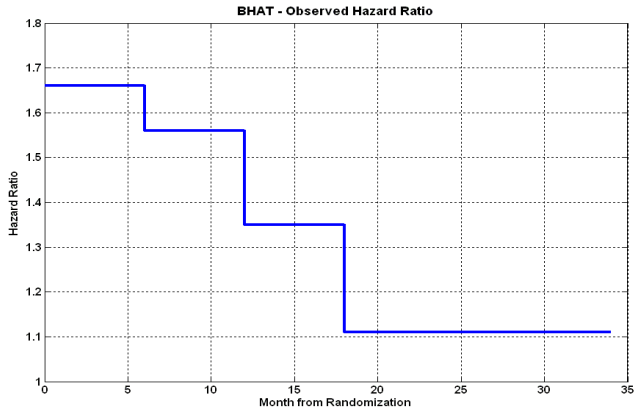
BHAT: Hazard Ratio

Slide given to me by Ed Lakatos



BHAT: analysis with different method





Kosorok and Lin (JASA 1999): Logrank was specified prior to initiating the trial, but with the weighted $G^{20,0}$ logrank, the trial could have perhaps been stopped 10 months earlier during which 36 treatment and 58 placebo deaths occurred

Authors make point for illustration only.
Also, weighted tests were new when trial done, but still…

NEJM (2015, Heeney et al.)

A Multinational Trial of Prasugrel for Sickle Cell Vaso-Occlusive Events

- Inherited blood disorder characterized by painful crises
- Placebo-controlled, patients 2-17 years, 9-24 months treatment,
  N = 341, sites in Americas, Europe, Asia, Africa.
- **Primary EP**: # vaso-occlusive crises (VOC). At least 2 VOCs pre-study
  Secondary EPs: sickle cell-related pain, pain intensity. Daily diary.

Analysis method: Anderson-Gill (adjusted for hydroxyurea use, age group)

Sickle Cell Trial

- What is the best primary endpoint?
  - # of VOCs? Time to first VOC event?
  - Diary reported pain?

- Given # of VOCs as primary endpoint, what is the best method of analysis?
  - Anderson-Gill (this trial), MH row means score (different trial*), negative binomial regression?
  - When these tests are applied to the same data, results can be different

*British Journal of Hematology (2011, Ataga et al.)

Efficacy and Safety of Epoetin Alfa in Critically Ill Patients

NEJM (2007, Corwin et al.)

- Anemia in critically ill is treated with red-cell transfusions
  Hypothesis: epoetin alpha may reduce need for red-cell transfusions

- Trial enrolled 1460 medical, surgical, trauma patients 48-96 hours after ICU admission. Epoetin alpha or placebo administered for maximum of 3 weeks. Patients followed for 140 days

- **Primary EP**: whether patient received red-cell transfusion
  Secondary EPs: no. of red-cell units transfused, mortality and change from baseline in hemoglobin

Anemia in Critically Ill Trial

Mortality by Day 140

Day 140	Epoetin alpha	Placebo	Hazard Ratio (95% CI) Unadjusted	Adjusted*
All patients	104/733 (14.2)	122/727 (16.8)	0.83 (0.64–1.08)	0.86 (0.65–1.13)
Admission group				
Trauma	24/402 (6.0)	36/391 (9.2)	0.63 (0.38–1.06)	0.40 (0.23–0.69)
Surgical, nontrauma	27/162 (16.7)	27/168 (16.1)	1.02 (0.60–1.74)	0.91 (0.52–1.60)
Medical, nontrauma	53/169 (31.4)	59/168 (35.1)	0.88 (0.60–1.27)	0.99 (0.66–1.49)

*Adjusted for age, sex, admission group, APACHE II score, baseline hemoglobin, iron, serum creatinine, types of co-existing conditions, injury severity score

CONCLUSIONS

The use of epoetin alfa does not reduce the incidence of red-cell transfusion among critically ill patients, but it may reduce mortality in patients with trauma. Treatment

Controlled Clinical Trials

Volume 18, Issue 6, December 1997, Pages 550–556

Eighth International Symposium on Long-Term Clinical Trials

Secondary endpoints cannot be validly analyzed if the primary endpoint does not demonstrate clear statistical significance

Robert T. O'Neill, PhD

Office of Epidemiology and Biostatistics, Center for Drug Evaluation and Research/FDA, Rockville, Maryland, U.S.A.

Activities Before, During and After Phase 3 Trials

- Dose selection, protocol development, endpoint selection, EOP2 meeting, label discussions, choice of control group, competitive positioning, adjudication committee, global site selection, patient I/E criteria, site monitoring, investigator meetings, DMCs, boundaries for group sequential trials, safety issues, SAPs, TFL specs, US-EU requirements, protocol deviations, edit checks, blinded data reviews, SAS programming, validation, NDA / BLA preparation

- Immense effort from protocol development to **FPFV** to **LPLV** to **unblinded analysis.** The time from start to finish is several years. Hopes are raised. Cost can exceed $100MM.

- Yet we let inference depend on a single statistic

Proposal

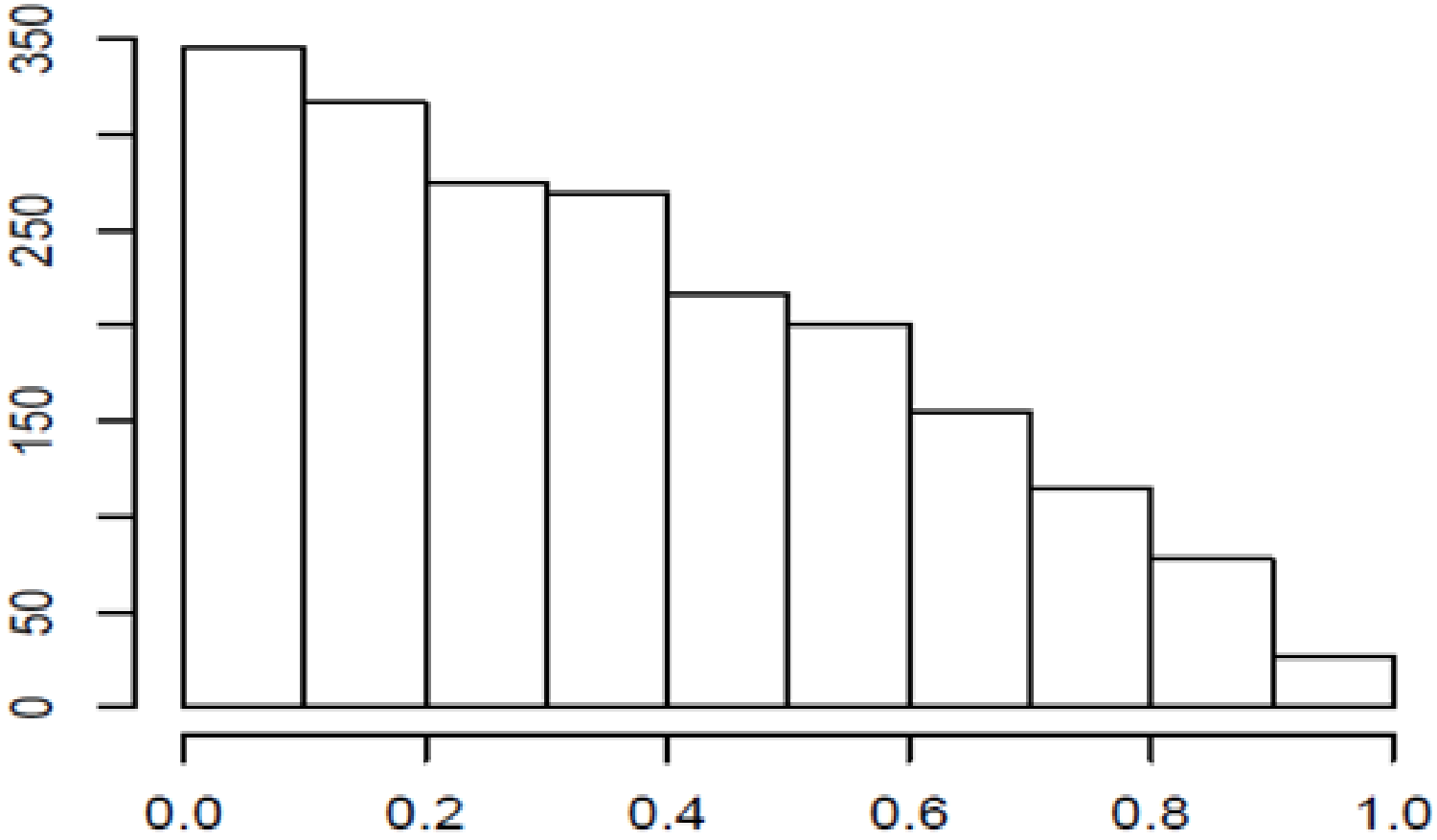
- We need to continue to pre-specify
- But pre-specify more than one test for formal inference
  - Pre-specify k >1 tests for formal inference. Get p-values

Combine p-values via some function. We'll look at two:
(a) min of p-values, minP and (b) Fisher's combination test

$$FCT = -2\sum \log(p_i)$$

Ganju, Yu and Ma (Pharm Stat 2013) and Ganju and Ma (Stat Method Med Res 2014)

Maintaining the Type I Error Rate



Frequency Histogram of minP

P-value of the Combined Test

- Null hypothesis: the two treatment groups are identical

- Use permutations. Treatment labels are re-assigned at random and test statistic calculated. Repeat multiple times. This gives the null distribution from which the p-value of a combination test is determined

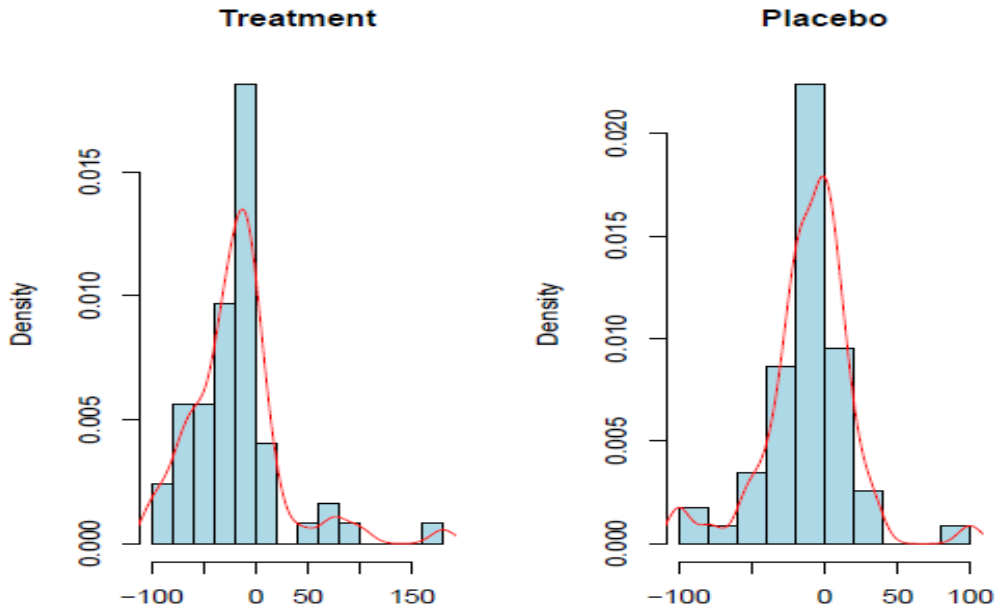
  For example, the p-value of minP, p_{minP}, is the proportion of permutation-generated minPs that are less than the observed minP

Let’s see how the method works in practice

Example 1: Parametric, Non-parametric

- Placebo-controlled trial in patients with rheumatoid arthritis
- Endpoint is % improvement from baseline in the Disability Index of Health Assessment Questionnaire

  N/group ≈ 60
- Covariates: AGE65, SEX



Results

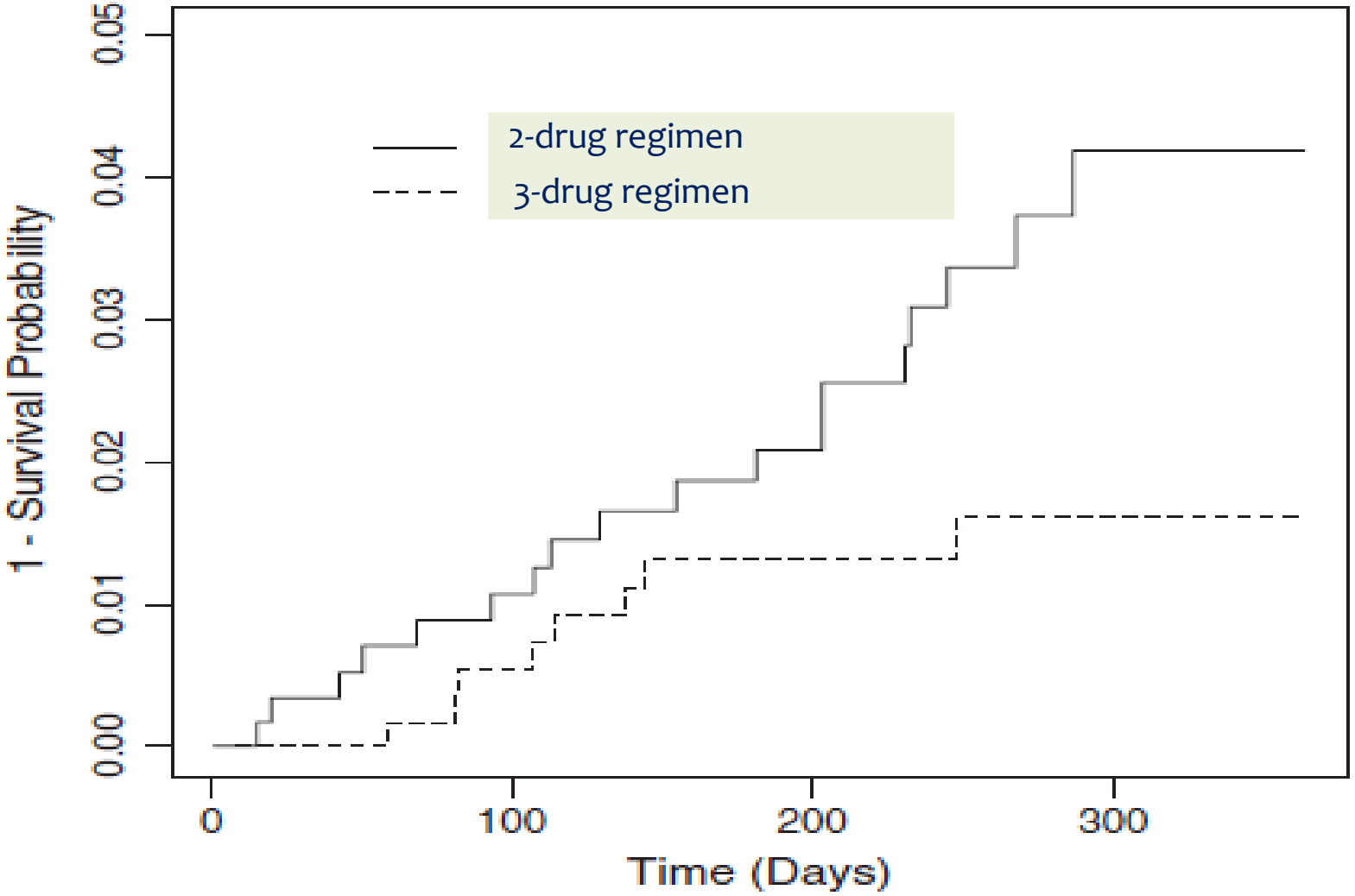
	P-values
• T-test	0.130
• Adjusted for age	0.120
• Adjusted for sex	0.140
• Wilcoxon rank sum test	0.012
• minP	0.012
• p_{minP}	0.020

Example 2: Time to Event

- AIDS Clinical Group Study 320: Hosmer and Lemeshow (2008)
- Trial in HIV-infected patients
  - **3-drug regimen**: Indinavir, zidovudine or stavudine, lamivudine
  - **2-drug regimen**: zidovudine or stavudine, lamivudine
- Endpoint: time to death

  N/group ≈ 574
- Covariates: CD4 >/<= 50 cells/uL, age, sex

KM Curves


2-drug regimen
3-drug regimen
1 - Survival Probability
0.00
0.01
0.02
0.03
0.04
0.05
0
100
200
300
Time (Days)

Results

- Covariates in Cox regression models

	P-values
CD4	0.024
CD4, age, sex	0.019
p_{minP}	**0.018**

Example 3: Data Transformation

- Small data set (*Introduction to the bootstrap*, Efron and Tibshirani 1993, Ch 9)
- Endpoint is amount of remaining hormone (mg) in the devices sampled from 3 lots. X = no. of hours device worn.
- N = 9 / lot. Lots 1 and 2 compared.

	P- values
T-test, adjusted for X	0.0004
T-test, adjusted for log(X)	0.2274
p_{minP}	**0.001**

Why is minP Robust?

- Suppose 2 t-tests are pre-specified

  (y = trt and y = trt + x)

  with power = 60% and 95%. $\alpha = 0.05$.

- The test that has 95% power when $\alpha = 0.05$ also has high power when $\alpha = 0.025$, which is the minP critical value if the 2 tests are independent. T-test with 95% power has 91% power with $\alpha = 0.025$. **Power of minP between 91 and 95%.**

Related Work

Gastwirth JL. The use of maximum efficiency robust tests in combining contingency tables and survival analysis. *JASA* 1985. (Also JASA 1966).

Tarone RE. On the distribution of the maximum of the logrank statistic and the modified Wilcoxon statistic. *Biometrics,* 1981.

Lee JW. Some versatile tests based on the simultaneous use of weighted log-rank statistics. *Biometrics,* 1996.

Others: Fleming, Harrington, O'Sullivan (JASA 1987), Self (Biometrics 1991), Efron (Biometrika 1997)

Edwards (Stat Med 1999): fit models to blinded data. Use model selection algorithm to select best model. Fit that model to unblinded data. Very nice paper.

Ganju, Yu, Ma (Pharm Stat 2013) – explicit recommendation to avoid reliance on single test. Proposal more flexible than earlier work

Ganju, Ma (Stat Met Med Res 2014) – increase in power with combined tests ($N > k$, $N < k$)

Lin, Zhou, Ganju (J Biopharm Stats 2016) – simultaneous rejection of null in subgroups and overall sample

Ganju, Lin, Zhou (under review) – same point as 2013 paper but for GSTs

Should Logrank be Used as the Solo Test?

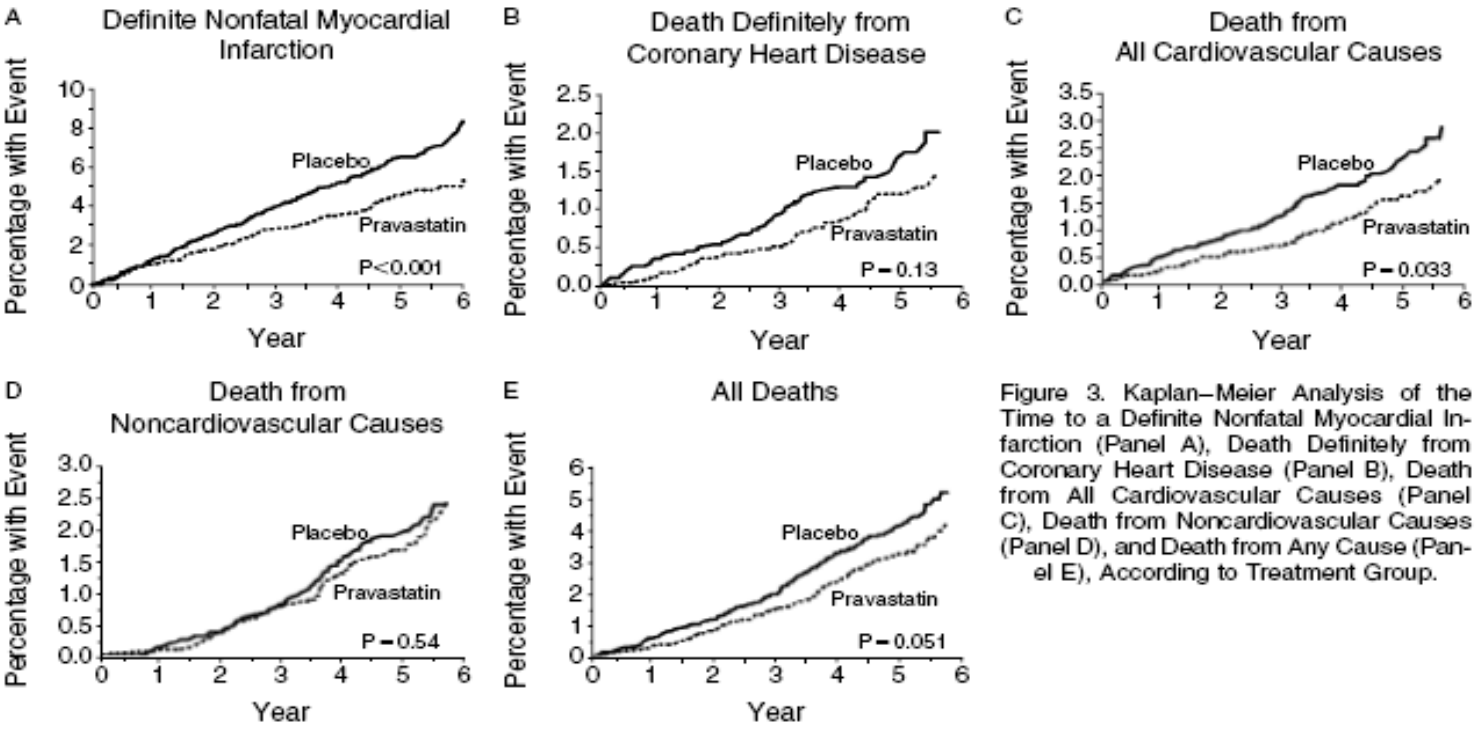


Figure 3. Kaplan–Meier Analysis of the Time to a Definite Nonfatal Myocardial Infarction (Panel A), Death Definitely from Coronary Heart Disease (Panel B), Death from All Cardiovascular Causes (Panel C), Death from Noncardiovascular Causes (Panel D), and Death from Any Cause (Panel E), According to Treatment Group.

PREVENTION OF CORONARY HEART DISEASE WITH PRAVASTATIN IN MEN WITH HYPERCHOLESTEROLEMIA

NEJM , Shepherd et al, 1995

Should Logrank be Used as the Solo Test?

Unadjusted and adjusted N from Lakatos (Biometrics, 1988)

- Logrank sample size for cardiovascular trial

  Uniform enrollment, power = 90%, α = 0.05, lag, average follow-up of 5 yrs, 2 yrs of recruitment

- N, no adjustment ~ 4400

  N, adjustment* ~ 8000

  *lag, non-compliance, drop-in

In practice, to maintain power N increased to allow for departures from PH because logrank is specified. With combined tests, trial size can be smaller.

MAIN PAPER

Pharmaceutical Statistics

(wileyonlinelibrary.com) DOI: 10.1002/pst.1582

Published online 6 August 2013 in Wiley Online Library

Robust inference from multiple test statistics via permutations: a better alternative to the single test statistic approach for randomized trials

Jitendra Ganju,[a]* Xinxin Yu,[b] and Guoguang (Julie) Ma[a]

The ubiquity of the logrank brings us to an important recommendation: use of the logrank as a single pre-specified test statistic should be discontinued unless there is a strong a priori reason for believing in proportional hazards. It should be replaced with analysis by multiple tests one of which could be the logrank. The

Extensions to Linear Models When

- Error df, N – K, small
- N < K

Start with N > K

- Set up: N = 20 and K = 16 binary covariates*

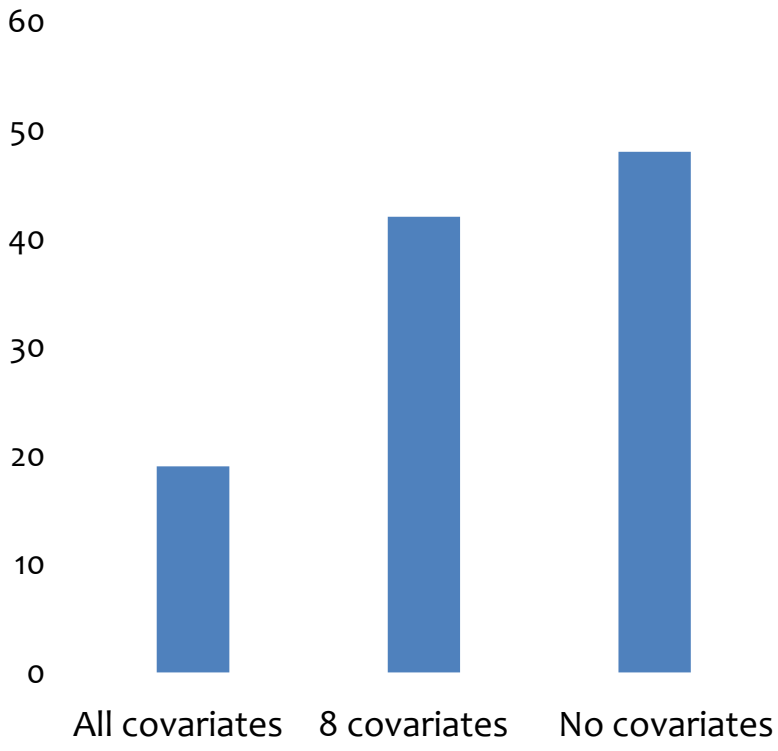
$$Y = a + b_o X_o + \sum b_k X_k + e$$

- What model to fit?

*Each covariate takes values 0 or 1 with equal probability. Unconditional expectation of estimator of treatment effect unbiased. e~N(0,1).

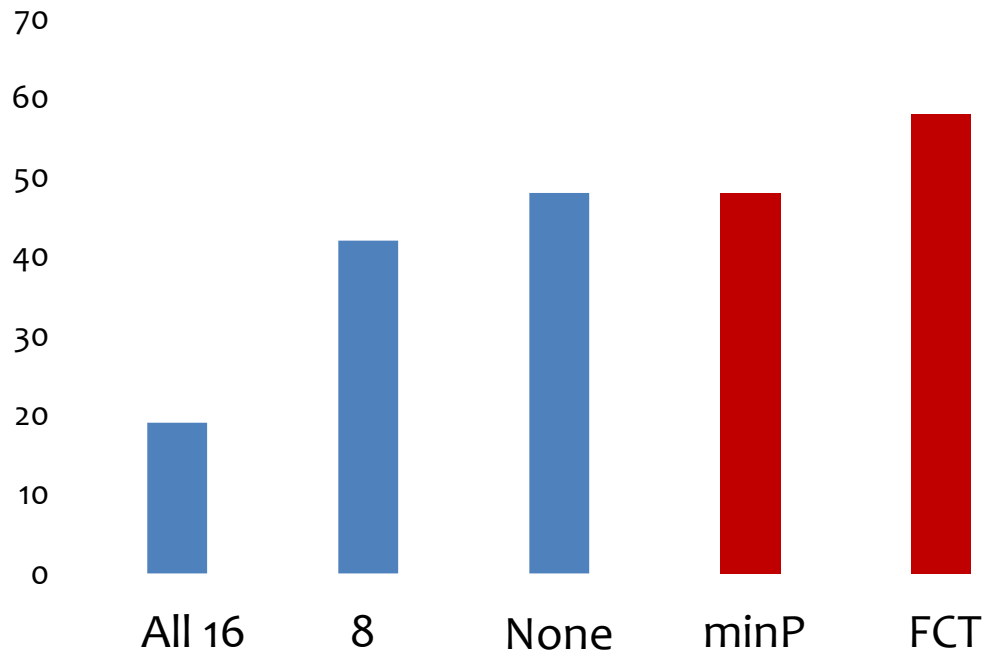
Power (%) when N = 20, K=16: $Y = a + 2X_0 + 1\sum X_K + e$

Power of a single test between 19% and 48%



Power (%) when N = 20, K=16: $Y = a + 2X_0 + 1\sum X_K + e$

minP and Fisher power from 2 t-tests: $T(X_1,\ldots X_8)$ and $T(X_9,\ldots,X_{16})$

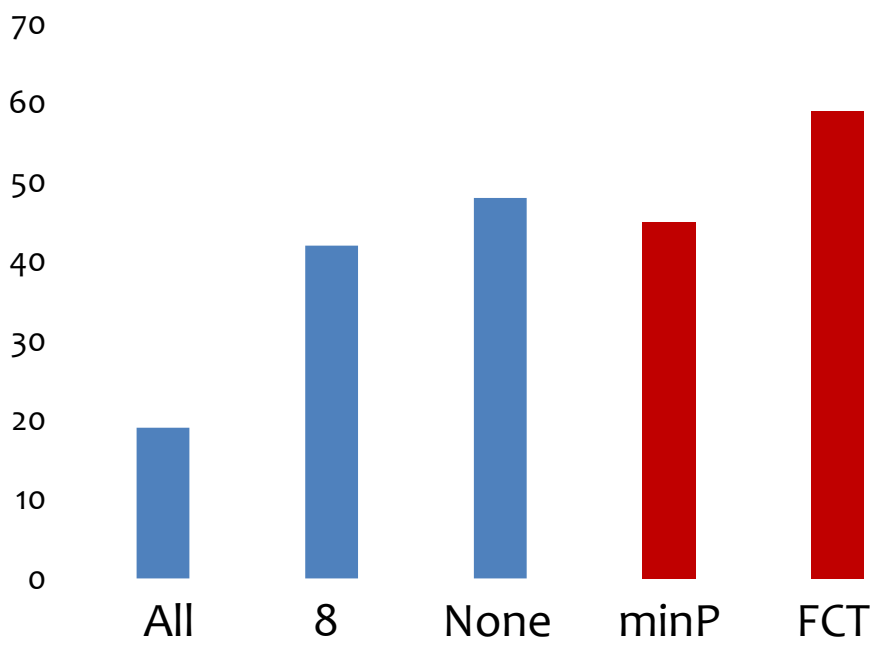


Combining tests can give more power than best single test

Power (%) when N = 20, K=16: $Y = a + 2X_o + 1\sum X_K + e$

minP and FCT power from **3** t-tests:
$T(X_1,\ldots,X_8)$, $T(X_9,\ldots,X_{16})$ and $T(X_1,\ldots,X_{16})$

Power from 2 tests nearly the same



Combining tests accommodates uncertainty and can give greater power

N < K

- Set up: N = 16 and K = 20 binary covariates

$$Y = a + b_o X_o + \sum b_k X_k + e$$

- Not possible for one model to include 20 covariates

Proposed approach, example

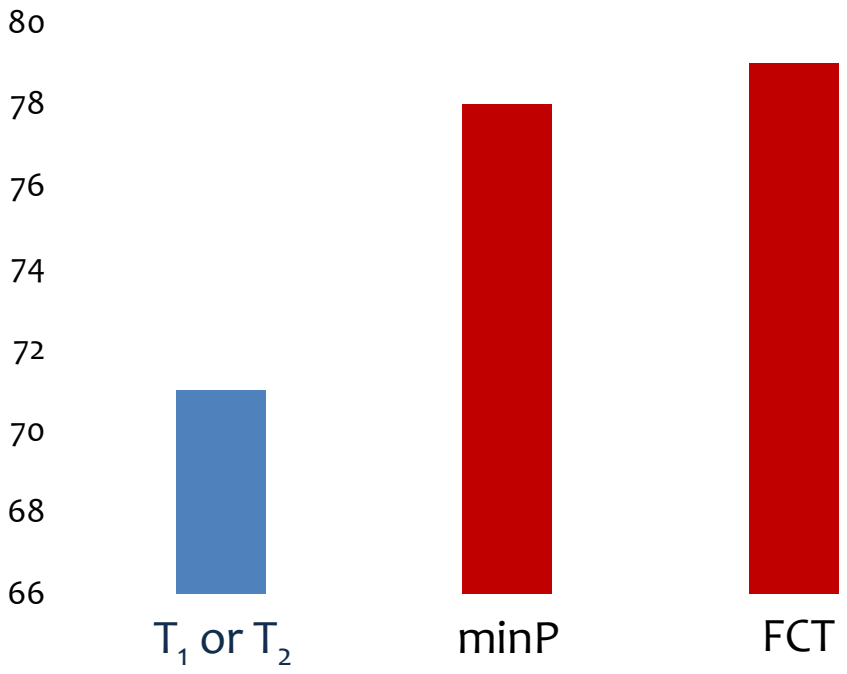
$Y = a + b_o X_o + \sum b_{K1} X_{k1}$ (k1 = 1 to 10)

$Y = a + b_o X_o + \sum b_{k1+1} X_{K1+1}$ (remaining 10)

Power (%) when N = 16, K=20: $Y = a + 4X_o + .5\sum X_K + e$

$T_1 = T(X_1,\ldots,X_{10})$ and $T_2 = T(X_{11},\ldots,X_{20})$

minP and FCT power from T_1 and T_2

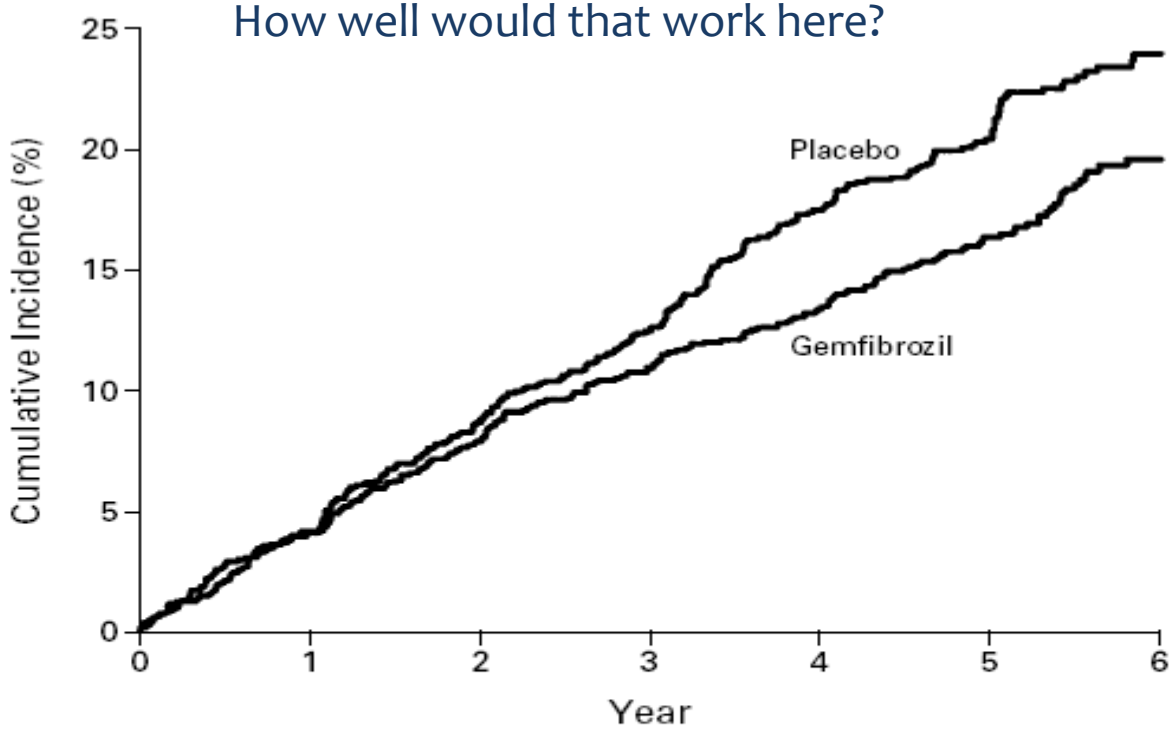


Extensions to…

Group Sequential Trials

Interim Looks: Practice is to Use the Same Test

How well would that work here?



No. at Risk							
Placebo	1267	1200	1118	1040	962	666	466
Gemfibrozil	1264	1201	1128	1067	1005	706	498

Figure 2. Kaplan–Meier Estimates of the Incidence of Death from Coronary Heart Disease and Nonfatal Myocardial Infarction in the Gemfibrozil and Placebo Groups.

The relative risk reduction was 22 percent (P=0.006), as derived from a Cox model.

Rubins et al.
NEJM 1995

Power in GSTs

$$\mathrm{T} = -\frac{ln(U)}{\lambda exp(X+Y)}$$

$\alpha_1 = 0.01$
$\alpha = 0.05$
$\mathrm{N} = 100$ at interim

$\mathrm{U} \sim U(0,1)$, $X \text{ and } Y \sim N(0,1), X \perp Y$, HR = 0.5, and $\mathrm{N} = 200$

Analysis	$Cox(X)$	$Cox(Y)$	$minP$	FCT
Interim	31			
Final	75			

Final power is conditional on not crossing the boundary at interim.

minP and FCT from Cox(X) and Cox(Y)

Power in GSTs

Analysis	$Cox(X)$	$Cox(Y)$	$minP$	FCT
Interim	31		31	38
Final	75		80	84

Power in GSTs: different tests at different times

$$T = -\frac{ln(U)}{\lambda exp(X+Y)}$$

$U \sim U(0,1)$, X and $Y \sim N(0,1), X \perp Y$, HR= 0.5, and $N = 200$

Analysis	Interim: $\boldsymbol{G^{5,0}}$ Final: $\boldsymbol{Cox(X)}$	$\boldsymbol{G^{0,0}}$ $\boldsymbol{Cox(Y)}$	$\boldsymbol{minP}(G^{5,0}, G^{0,0})$ $\boldsymbol{minP}(Cox(X), Cox(Y))$	$\boldsymbol{FCT}(G^{5,0}, G^{0,0})$ $\boldsymbol{FCT}(Cox(X), Cox(Y))$
Interim	12	20		
Final	78	75		

For single tests type I error rate ~ 6%. Can be corrected using permutations

Power in GSTs: different tests at different times

Analysis	Interim: $\boldsymbol{G^{5,0}}$ Final: $\boldsymbol{Cox(X)}$	$\boldsymbol{G^{0,0}}$ $\boldsymbol{Cox(Y)}$	$\boldsymbol{minP}(G^{5,0}, G^{0,0})$ $\boldsymbol{minP}(Cox(X), Cox(Y))$	$\boldsymbol{FCT}(G^{5,0}, G^{0,0})$ $\boldsymbol{FCT}(Cox(X), Cox(Y))$
Interim	12	20	18	21
Final	78	75	83	88

Power in GSTs: comment

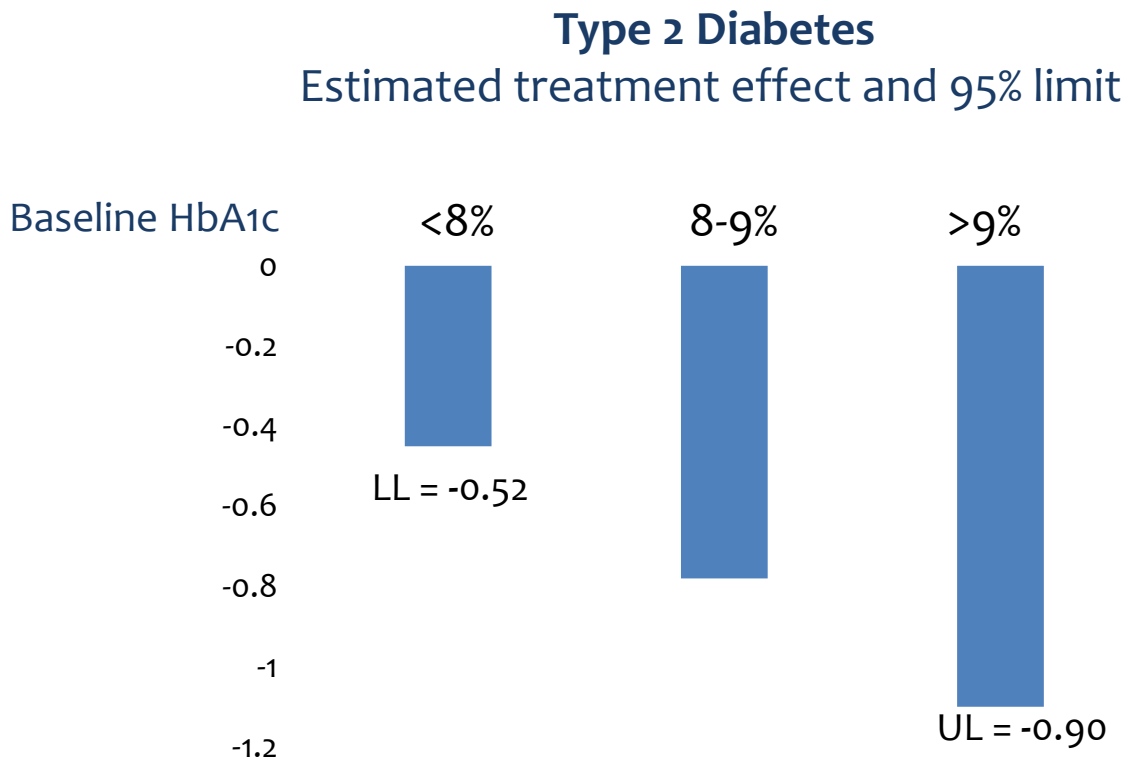
Even with single tests, no need to limit ourselves to the same test at each interim analysis time. Can use different single tests at each interim analysis.

Need to specify test for next interim before seeing results from current interim.

Extensions to…

Simultaneous rejection of the null in a subgroup and in the overall sample

Larger Effect in Patients with More Severe Disease



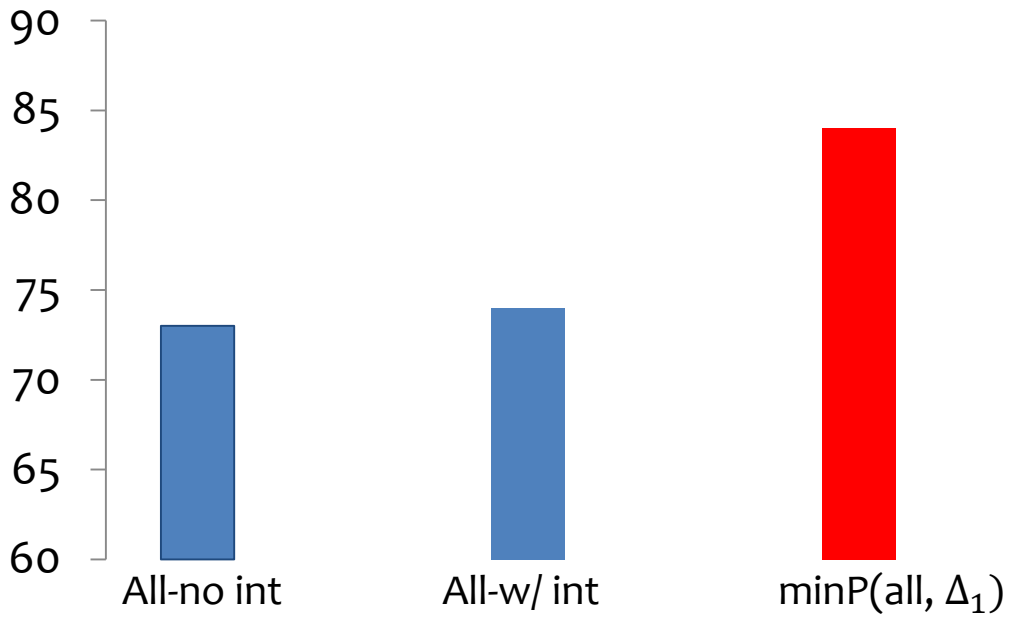
Source: Canagliflozin briefing book, FDA Ad comm, Fig 11 (visual est. here), Cana 100 mg - placebo

When we know effect larger for identifiable subgroups, propose method to use that fact. Conventional method cannot incorporate.

MinP Power from Subgroup and Full Model

N = 40/trt group

$\Delta_1 = 1$, $\Delta_2 = .2$, $\Delta = 0.6$, $\varepsilon \sim N(0, 1)$



Inference: if $p_{minP} \leq \alpha$, and $p(\Delta_1) \leq \alpha$, can reject null in subgroup and overall sample even if $p(\Delta) > \alpha$. Notes: (a) Need for careful interpretation. (b) Loss in power if wrong subgroup selected.

Future Work

- Example: Choose endpoints that capture different but important aspects of response to treatment

  e.g. time to first event, days hospitalized

- Normal way is to choose one. Let's look at it differently. Treatment efficacious if either EP demonstrates benefit. In addition, suppose 2 methods of analysis for each endpoint.
  - EP1: p_{11}, p_{12}, p_{minP1}
  - EP2: p_{21}, p_{22}, p_{minP2}

- Better for secondary endpoints

Future Work cont'd

- **Blinded data:** Use a selection criterion to order EPs. After unblinding proceed with combination tests on each EP.

- **Unblinded data:**

  1. Trial to make single statement if benefit exists*
     Single step: e.g. test based on p_{11}, p_{12}, and p_{21}, p_{22}

  2. Trial to make claim on any EP showing benefit
     Step-wise. Use with Hochberg method

  3. Non-inferiority trials

*O'Brien (Biometrics 1984), General: Dudoit, Shaffer, Boldrick (Statistical Science 2003)

Remarks

- The point is less on which combining function to use but on hedging our bets with combining different methods of analysis to reduce risk
- Power close to or better than best performing single test
- Sometimes we only have one shot. Emphasis should be on making inferences robust.

Back-ups

Example 4: Linear Model

- Data from SAS course book on mixed models (SAS 2009, Ch 7)

  "A pharmaceutical company compared the effects of two drugs, A and B, on a clinical measurement called *flush*… The original plan called for each drug to be randomly assigned to the same number of patients within each clinic."

- 10 centers (n ranged from 3 to 28), overall allocation of 68:83. Three missing observations are deleted

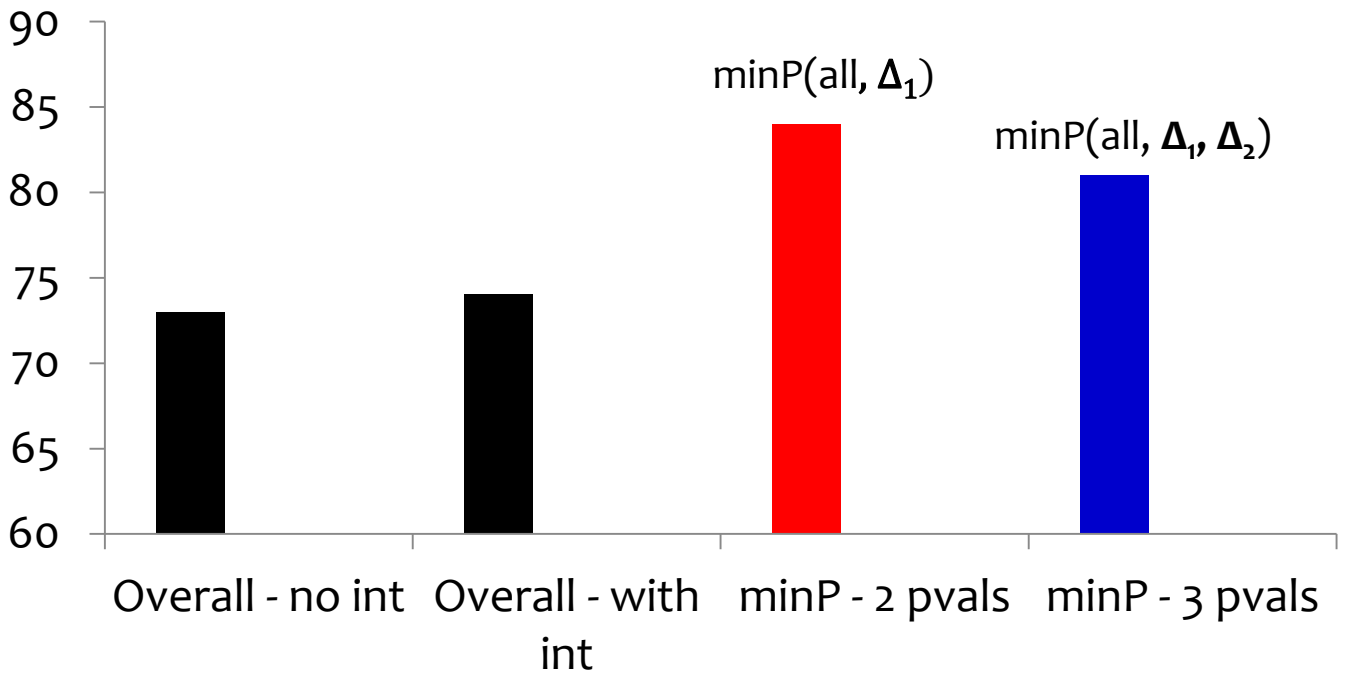
	p-values
Diff in means	0.024
Diff in mean adj for center	0.117
p_{minP}	**0.031**

Counting Each Subject Twice Can Give More Power Than Counting Once

$\Delta_1 = 1, \Delta_2 = .2, \Delta = 0.6$

minP based on subgroup1, all (red)

minP based on subgroup1, subgroup 2, all (blue)



Time to Event Trials: LR, wLR, minP

- Comparison of minP with logrank and weighted LR

Two weighted logrank (Harrington-Fleming) tests:

LR_E: $\hat{S}(t)$ more weight to early differences

LR_L: $1-\hat{S}(t)$ more weight to late differences

minP = min(p_{LR}, p_{LRE}, p_{LRL})

Power

Case	N	HR	Treatment	Control
1	200	Constant	0.07	0.10
2	200	Late diff	0.12	(0.12, 0.08, 0.06), t = (0-3, 3-6, 6-30)
3	200	Early diff	0.06	(0.12, 0.08, 0.06), t = (0-3, 3-6, 6-30)