# Heterogeneity Over Time in Clinical Trials

Keaven M. Anderson, Jason B. Clark, Kenneth S. Liu

Merck Research Laboratories

*EMEA / EFPIA Workshop on Adaptive Designs*

December 14, 2007

# Overview

- Focus on examples of heterogeneity in Phase III trials
- Raise questions related to
  - Heterogeneity due to learning in pivotal trials
  - Estimation issues in adaptive design
  - The challenge of distinguishing random and systematic heterogeneity

# Adaptation and Heterogeneity: Examples

- Changes in study practice (learning)
  - Cardiovascular disease example
- Adaptation and estimation bias
  - Oncology example: dose-selection
- “Random” variability in subsets
  - Cardiovascular disease example of sample size re-estimation
- “Systematic” enrollment/efficacy trends
  - Depression example

# EPIC Trial: Learning

- First large trial with potent anti-platelet agent abciximab (EPIC Investigators, 1994)
- Based on unblinded interim analysis of 700 of 2100 planned patients
  - DMC raised safety concern (bleeding)
  - Too few endpoints to assess efficacy
- Patient treatment guidance added by blinded steering committee
- Potential issues
  - Did patient selection change to avoid patients with high risk of bleeding?
  - Did this affect efficacy of each treatment group?

# EPIC Trial: Confirming

- Composite endpoint significant at end of trial
- Drug approved based on a single pivotal trial
  - Irreversible endpoint benefit was key to regulatory acceptance
  - Initially recommended only for high-risk patients
  - Use of drug was relatively limited
- Further studies expanded usage
- Question:
  - *Are there cases of adaptive studies where heterogeneity related to 'learning' may be tolerated, possibly requiring post-approval confirmation commitments?*

# Phase II/III Oncology Trial

- Two possible doses versus control
  - Early dose-selection followed by large confirmatory stage
- Dual primary endpoints
  - Early selection, futility and efficacy decisions based on progression-free survival (PFS)
  - Later efficacy confirmation and futility based on overall survival (OS)

# Phase II/III Oncology Trial

- Heterogeneity is ‘designed-in’ the trial
  - Interim analysis objectives
    - Interim 1 & 2
      - Limit patient exposure before strong proof of concept
      - Bounds designed to be informative
    - Interim 3
      - Possible accelerated approval based on PFS in US
      - Possible trial stop for positive survival result
    - Final analysis
      - Final confirmation of survival benefit

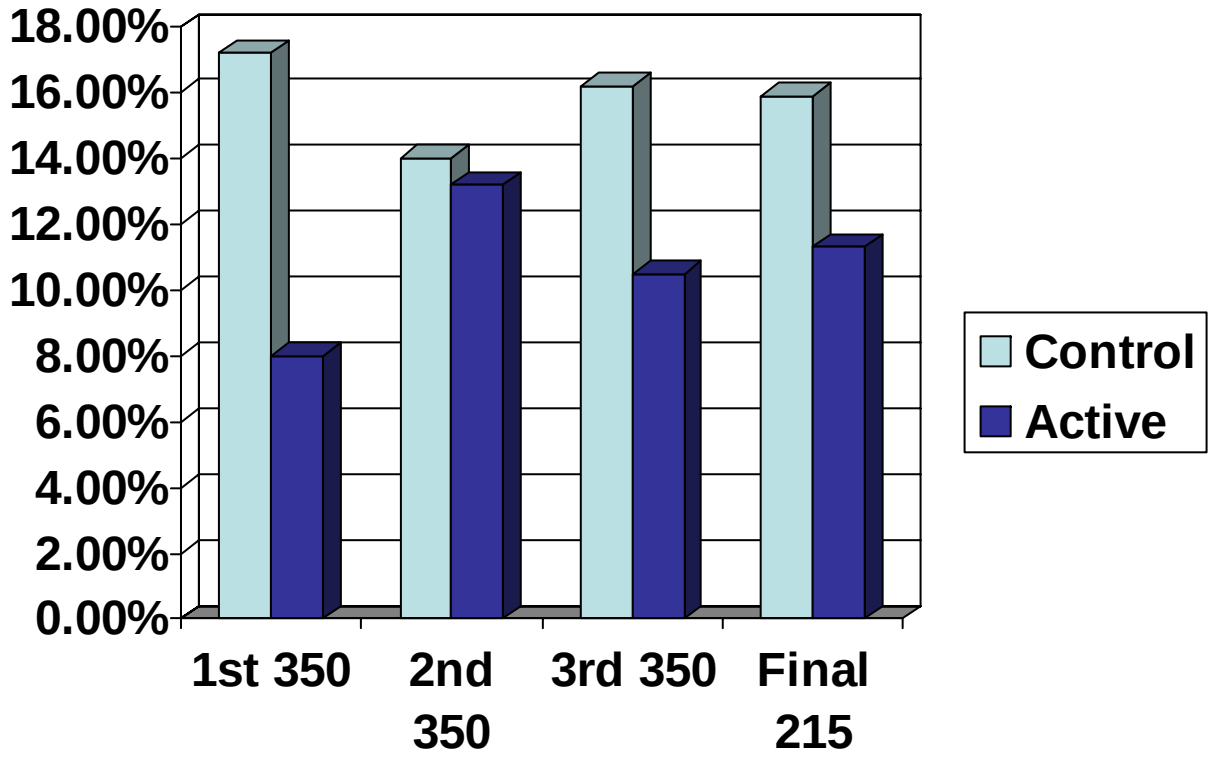
# Phase II/III Oncology Trial

- Dose selection and futility bounds introduce upward bias in naïve treatment estimates
  - Type I error well-controlled
  - Treatment benefit measured in 2nd, much larger part of trial may be a useful, unbiased, supportive analysis;
    - A confidence interval for treatment benefit in this subgroup may not be consistent with the overall test of significance
- Question:
  - *How much do estimation issues due to designed-in heterogeneity in adaptive designs raise concerns?*

# Random Variability

- The CAPTURE (CAPTURE Investigators, 1997) trial was a 1400, 2-arm trial in patients with unstable angina
- Composite endpoint: death, MI, urgent intervention, recurrent ischemia
- Interim analyses at 350, 700, 1050 patients
- Trial stopped for efficacy at 1050 patient analysis
- 215 patient over-run for 1265 total patients

# CAPTURE Trial: Sequential Sets of 350 Patients




**P=0.37 for heterogeneity**

# CAPTURE Example Conclusions

- Heterogeneity seemed substantial, but was consistent with random variation
- However, more sites and countries contributed after enrollment of first 350 patients
  - You can probably always find a *post hoc* explanation for heterogeneity as seen in the first 350
- Potential rationale for adaptation would be that over-enrollment of 215 patients could have been avoided after interim that stopped trial
  - Simulations confirm this is a competitive strategy if adaptation delayed until 700 patient interim

# Depression Example: "Systematic" Heterogeneity

- Hypothesis
  - 'Best' patients waiting to be enrolled at beginning of trial
  - Exhausted patient pool or rush to enroll at end of trial may produce less suitable patients
  - **More benefit may be observed in patients enrolled early versus late**
- Following analysis evaluates the above (Liu, et al, 2007):
  - Active control groups using paroxetine compared to placebo in 4 trials
  - Patients divided into quartiles of entry time within each trial; quartiles combined across trials
  - **Analysis of patient response by quartile of enrollment**
- Outcome
  - **Less benefit was seen among the last patients enrolled compared to earlier patients**
  - **Is this systematic heterogeneity or just another example of random variation?**

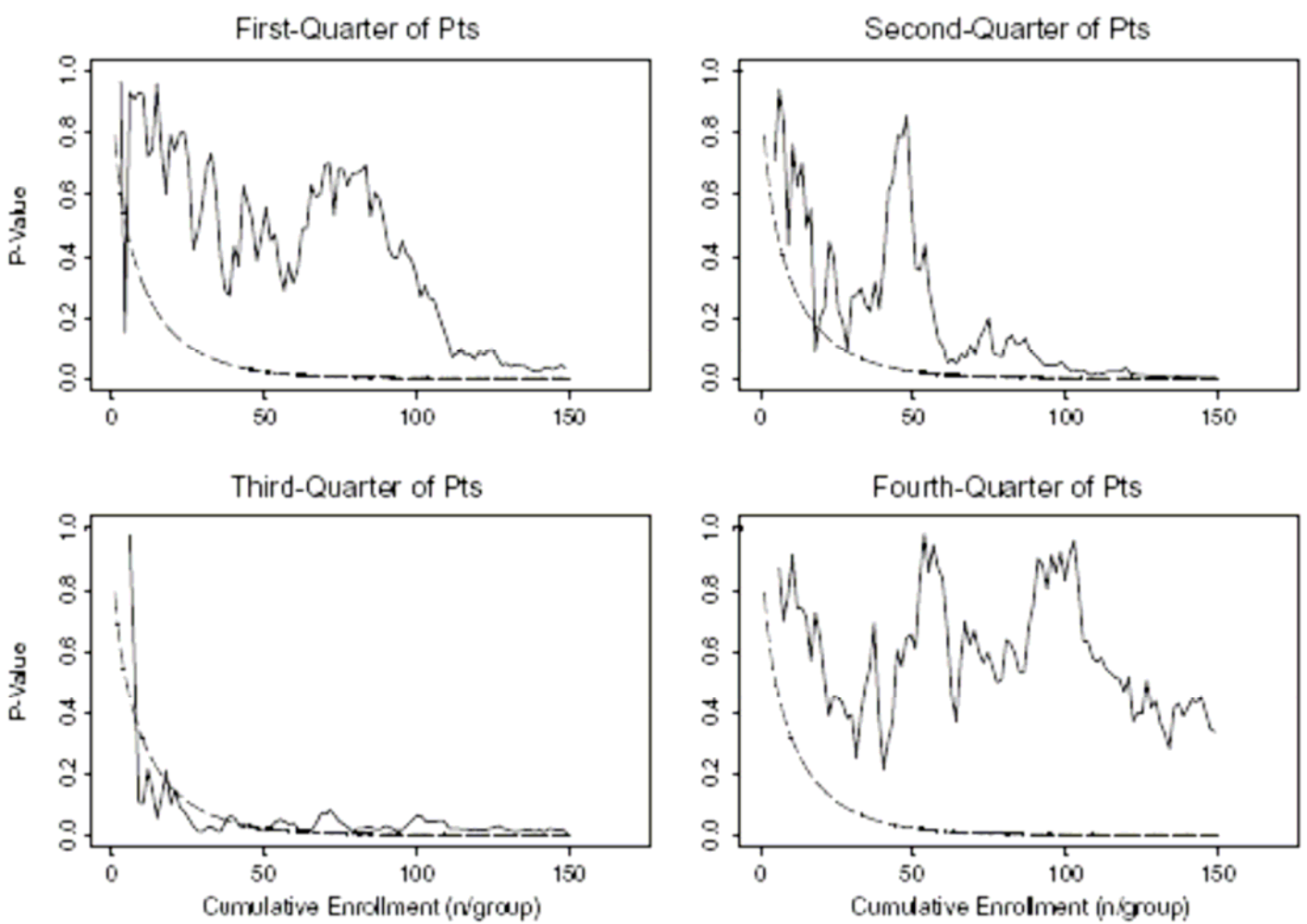
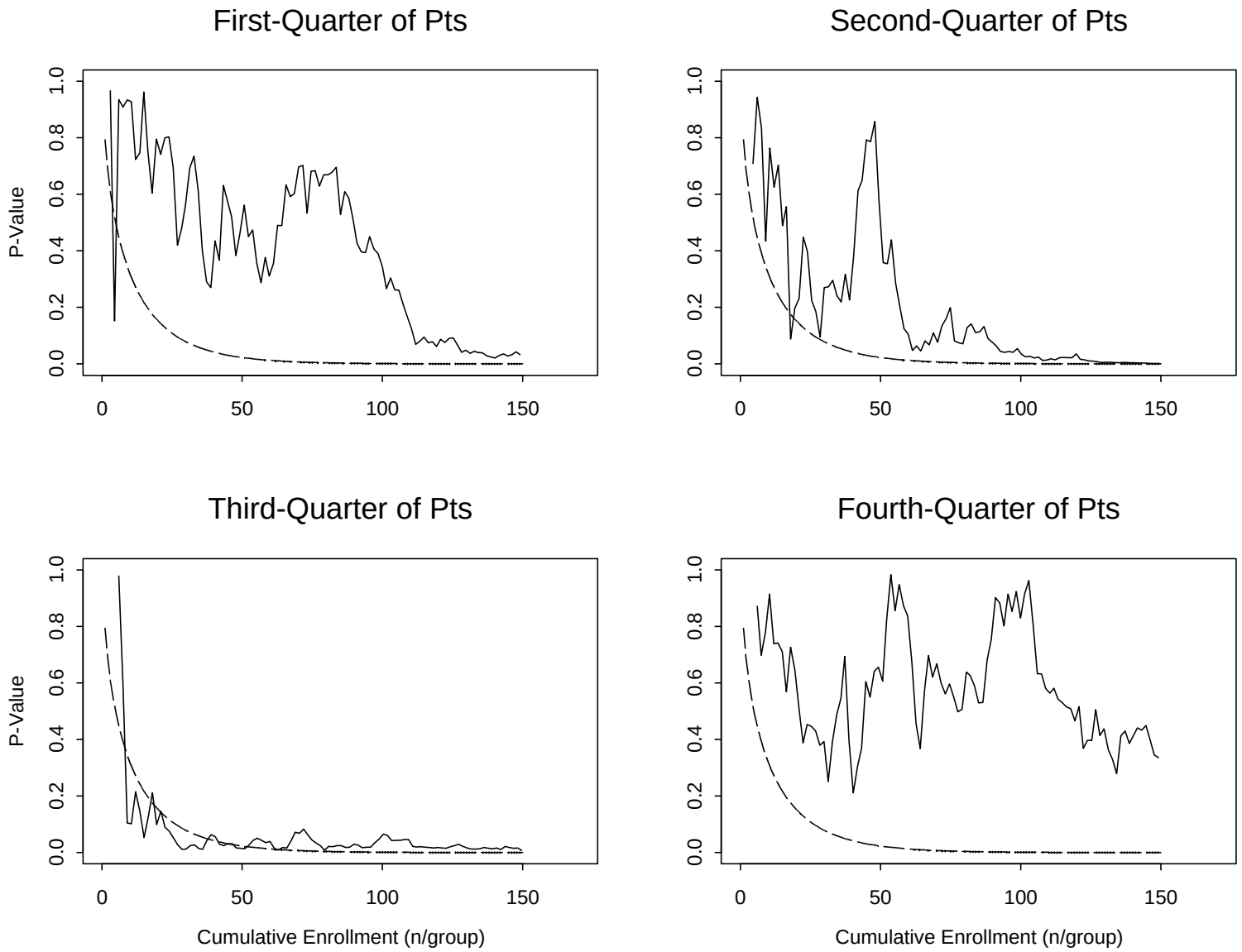



Fig. 4. *p*-Value for the treatment difference between paroxetine and placebo (in mean change from baseline HAM-$D_{17}$ score at week 8) versus cumulative enrollment by patient quarters using pooled data from four depression trials. The solid curve is the observed *p*-value. The dashed curve is the expected *p*-value for a standardized effect size of 0.3 versus cumulative enrollment.

# P-Values for accumulating data within quartile of enrollment time



# Discussion Points

- Concern: random or systematic heterogeneity can make it difficult to adapt appropriately
  - But inference concerning the global null hypothesis should not be an issue
- Heterogeneity due to changes in patient selection and care may occur during a trial for many reasons other than design adaptation:
  - Safety findings
  - Expanding sites/regions enrolling
  - Site experience
- Heterogeneity has often not prohibited trials from being confirmatory/approvable

# Questions

- Are there cases of adaptive studies where heterogeneity related to ‘learning’ may be tolerated, possibly requiring post-approval confirmation commitments?
- How much do estimation issues due to designed-in heterogeneity in adaptive designs raise concerns?
- Given that heterogeneity in treatment effect among sequential subgroups can be large due to random variation, would the expectation of homogeneity in an adaptive trial create a double-standard compared to other designs?

# References


- The EPIC Investigators (1994). Use of a monoclonal antibody directed against the platelet glycoprotein IIb/IIIa receptor in high-risk coronary angioplasty. *The New England Journal of Medicine*, 300:956-961.
- The CAPTURE Investigators (1997). Randomised placebo-controlled trial of abciximab before and during coronary intervention in refractory unstable angina: the CAPTURE Study. *Lancet*, 349:1429-35.
- Liu, KS, Snavely, DB, Ball, WA, Lines, CR, Reines, SA and Potter, WZ (2007). Is bigger better for depression trials? *Journal of Psychiatric Research*, in press, doi:10.1016/j.jpsychires.2007.07.003