Comparing Conditional and Predictive Power to Assess Futility in a Phase III Program with Two Studies

Greg Cicconetti, Ph.D.
GlaxoSmithKline

BASS XVI
November 9, 2009

Outline

- Phase III program characteristics
- Overview of futility methods considered
- Review of Bayesian model for hazard ratio based on normal distributions
- Discussion of conditional power and predictive power with reference of this phase III program

Phase III Program Characteristics

- Two large, similarly designed, event-driven trials in *low risk* and *high risk* patients are intended to support registration

- Common features of both trials
  - Both studies are Standard of Care (SOC) vs. SOC + Drug
  - Common primary and key secondary endpoints
  - Trials enroll sufficient subjects to collect their required events in <4 yrs
  - Common Hazard Ratio (HR)
  - Common dropout and IP discontinuation rates
  - Group sequential design (GSD) with 2 interim analyses (IAs) scheduled based on information rates

More Common Features

- 2 IAs + final conducted at 850, 1150, and 1500 events
  - Two sided test of H_0: HR = 1
  - Type I error control at 0.05 level, 90% power when HR = 0.845
  - Info Rates: 56.67%, 76.67%, 100% of required events.

- Small P-values required to stop early for efficacy
  - At interim 1: Stop and reject if P-value < 0.0005
  - At interim 2: Stop and reject if P-value < 0.001

- If trial continues to the planned number of events
  - At final: reject if P-value < 0.0499

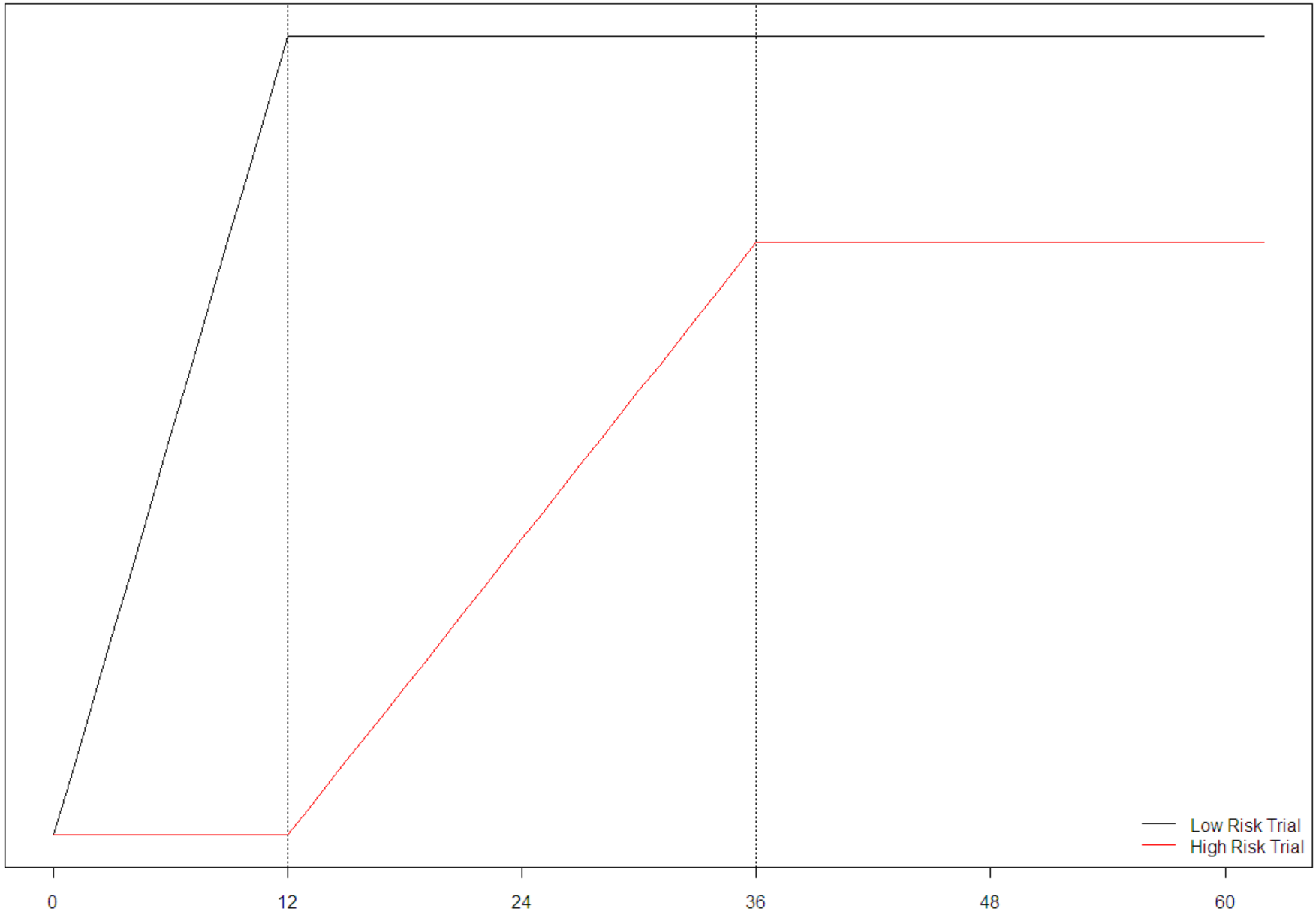
Phase III Program Characteristics

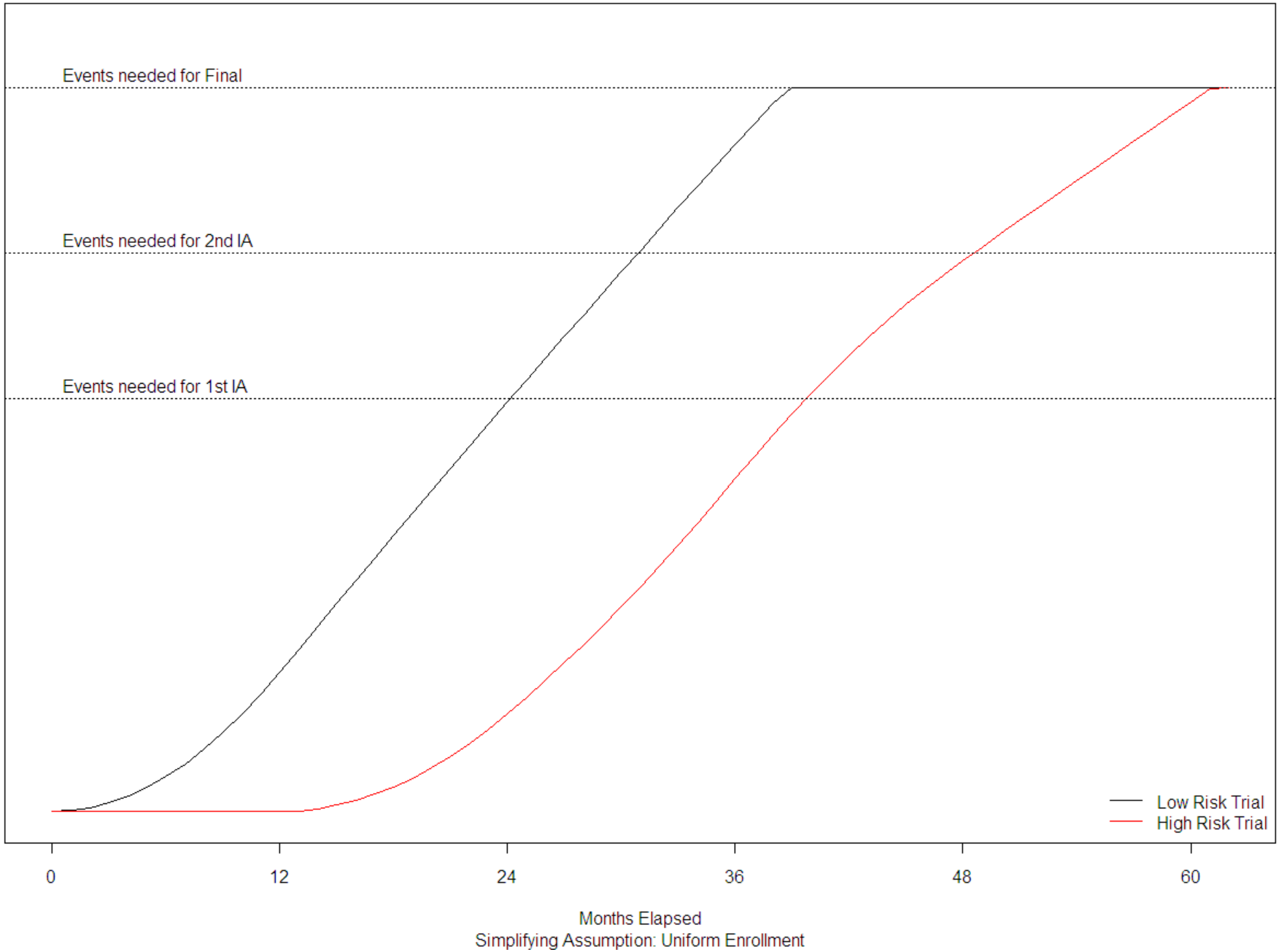
- Idiosyncrasies
  - Low Risk Trial (LRT) starts ahead of the High Risk Trial (HRT)
    - Pending outcome of a safety substudy in LRT, HRT starts ~1 yr later
  - LRT expected to enroll more quickly
  - HRT assumed to have to have a larger initial event rate which drifts towards the low risk population event rate over time
    - HRT time to event assumed to be piecewise exponential
  - HRT enrollment will be more challenging

Sample Sizes

- These design characteristics were used to explore the relationship between study duration and sample size
  - Both trials enroll 11k+ subjects
  - LRT enrolls ~4k more patients

- HRT is rate limiting for completing the phase III program


Patient Accrual in Phase III Program
Low Risk Trial
High Risk Trial
0
12
24
36
48
60
Months Elapsed
Simplifying Assumption: Uniform Enrollment


Event Accrual in Phase III Program
Events needed for Final
Events needed for 2nd IA
Events needed for 1st IA
Low Risk Trial
High Risk Trial
0
12
24
36
48
60
Months Elapsed
Simplifying Assumption: Uniform Enrollment

Independent Data Monitoring Committee

- IDMC has discretion to add interims analyses or recommend a trial continue/stop
  - Alpha spending function to be used as needed
- Reviews safety data biannually

- A pragmatic philosophy is desired
  - Group sequential boundaries are non-prescriptive guidelines for stopping due to efficacy
    - While the statistical methods are often very useful, the ultimate recommendation to terminate or continue depends largely on the judgment of a data monitoring committee and the initial guidance provided by the trial steering committee of investigators and sponsors. – DeMets
  - Direction of departure from null, trends of treatment effect over time, cost of continuing, all play a role

Overview of Futility Monitoring Options Considered

- Deterministic Methods

- Conditional Power
- Predictive Power

- Related ideas
  - Posterior probabilities regarding underlying treatment effect
  - Predictive probabilities regarding observed treatment effect

Motivation for Current Discussion

- Given the alpha spending, our expectation is that the LRT runs to completion
  - IDMC is wary about futility testing at 1st IA (immature data)
  - At 2nd IA, momentum expected to push the study to completion, even if drug is thought to be futile
    - More scientific value in having a completed study
    - Events require adjudication
    - Not cost effective to stop

- What about the HRT?
- Could the data collected in the LRT assist us in making an informed decision about futility in the HRT?

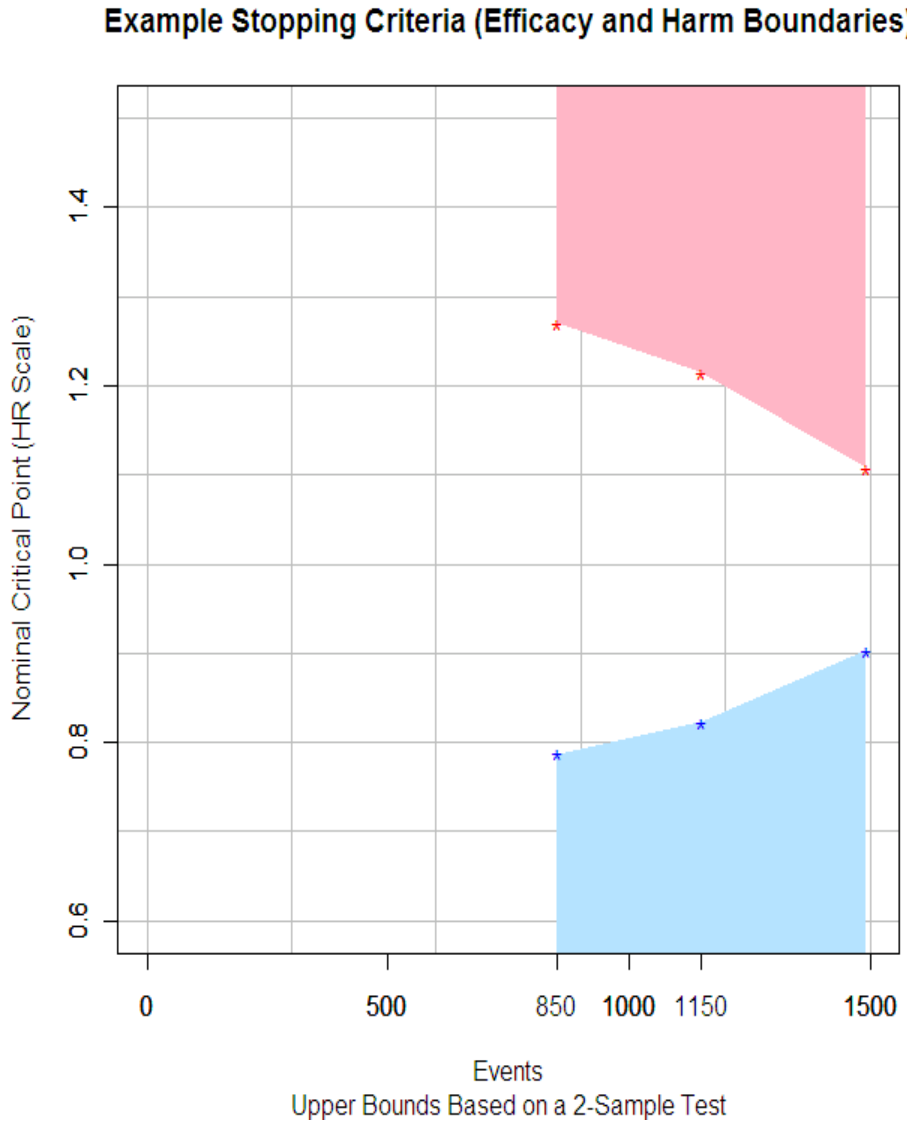
Deterministic Methods for Assessing Futility

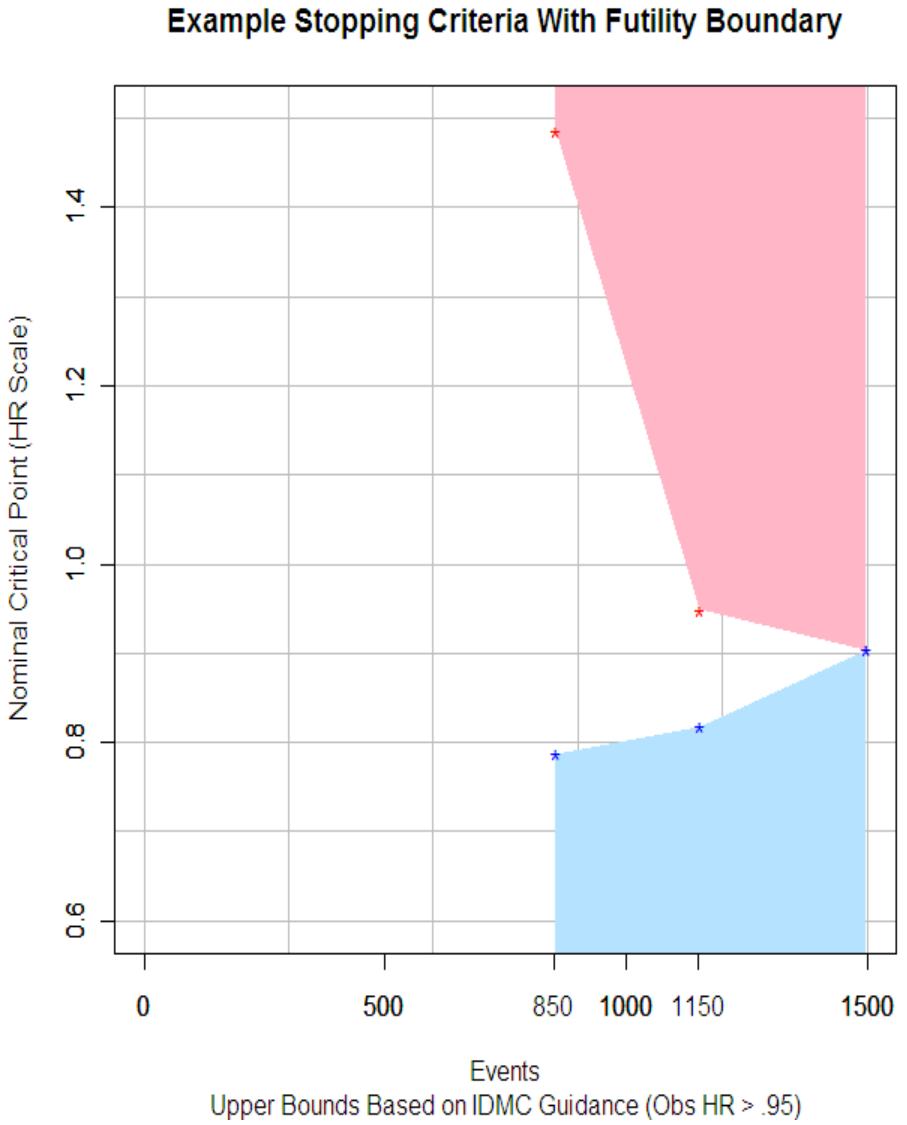
- Decision-making is tied to currently available data
  - Based on formal tests for futility and efficacy
  - Stopping boundaries based on how we propose to spend Type I and Type II error
  - Requires committing to pre-specified decision rules.

- No consideration for the impact of future observations
- No incorporation of external information
- May be inappropriate at early analyses or if treatment effect is delayed

- For this program, consensus is that **these methods do not suit us**

Example of Deterministic Stopping Rules





Stochastic Stopping Rules for Futility

- Decision-making is tied to predicting the outcome of study
  - Do we reject H_0: HR = 1 in the end?
- **Power**
  - P(Statistically significant observed HR in the end)
- **Conditional Power**
  - P(Statistically significant observed HR in the end **given interim data**)
  - Incorporates only within study information
  - Basic Conditional Power: Predictions based on hypothesized treatment effect
  - Adaptive Conditional Power: Predictions based on study estimate of treatment effect
- **Predictive Power**
  - Incorporates study data + prior info to make predictions
  - P(Statistically sig observed HR in end **given interim data + prior**)

Related Ideas

- These tools focus attention on *underlying* and *observed treatment effect*, respectively.

- **Posterior Probabilities**
  - Uses **study data + prior info** to describe the current thinking regarding the *underlying treatment effect*
- **Predictive Probabilities**
  - Uses **study data + prior** info to make predictions regarding the study-end *observed treatment effect*
  - Predictive Probability statements could address

Pred.Prob($HR_{obs} < 0.9$)	= Pred.Prob (Drug is marketable)
Pred.Prob($HR_{obs} < 0.845$)	= Pred.Prob (Drug is effective)
Pred.Prob($HR_{obs} < 1$)	= Pred.Prob (Drug has some benefit)
Pred.Prob($HR_{obs} > 1$)	= Pred.Prob (Drug is detrimental)
Pred.Prob($0.97 < HR_{obs} < 1.03$)	= Pred.Prob (Drug is similar to comparator)

Modeling log(HR) with Normal Distributions

- Work with treatment effect on log scale: $\theta = \log(HR)$
- Normal Distributions used throughout: N(Mean, Variance)
  - **Prior distribution** $\sim N(\theta_{prior}, \ 4/m_0)$
  - **Estimates of HR given data** $\sim N(\theta_{obs}, \ 4/m)$
  - **Posterior Distribution** $\sim N(\theta_{post}, \ \Sigma_{post})$
  - **Predictive Distribution** $\sim N(\theta_{pred}, \ \Sigma_{pred})$

θ_{prior} = 'best guess for θ' (worth m_0 events)

θ_{obs} = estimate from data *based on m events*

$\theta_{post} = (m_0 \theta_{prior} + m\theta_{obs})/(m_0 + m)$

$\theta_{pred} = (m_0 \theta_{prior} + m\theta_{obs})/(m_0 + m)$

$\Sigma_{post} = 4(m_0 + m)^{-1} = ((4/m_0)^{-1} + (4/m)^{-1})^{-1}$

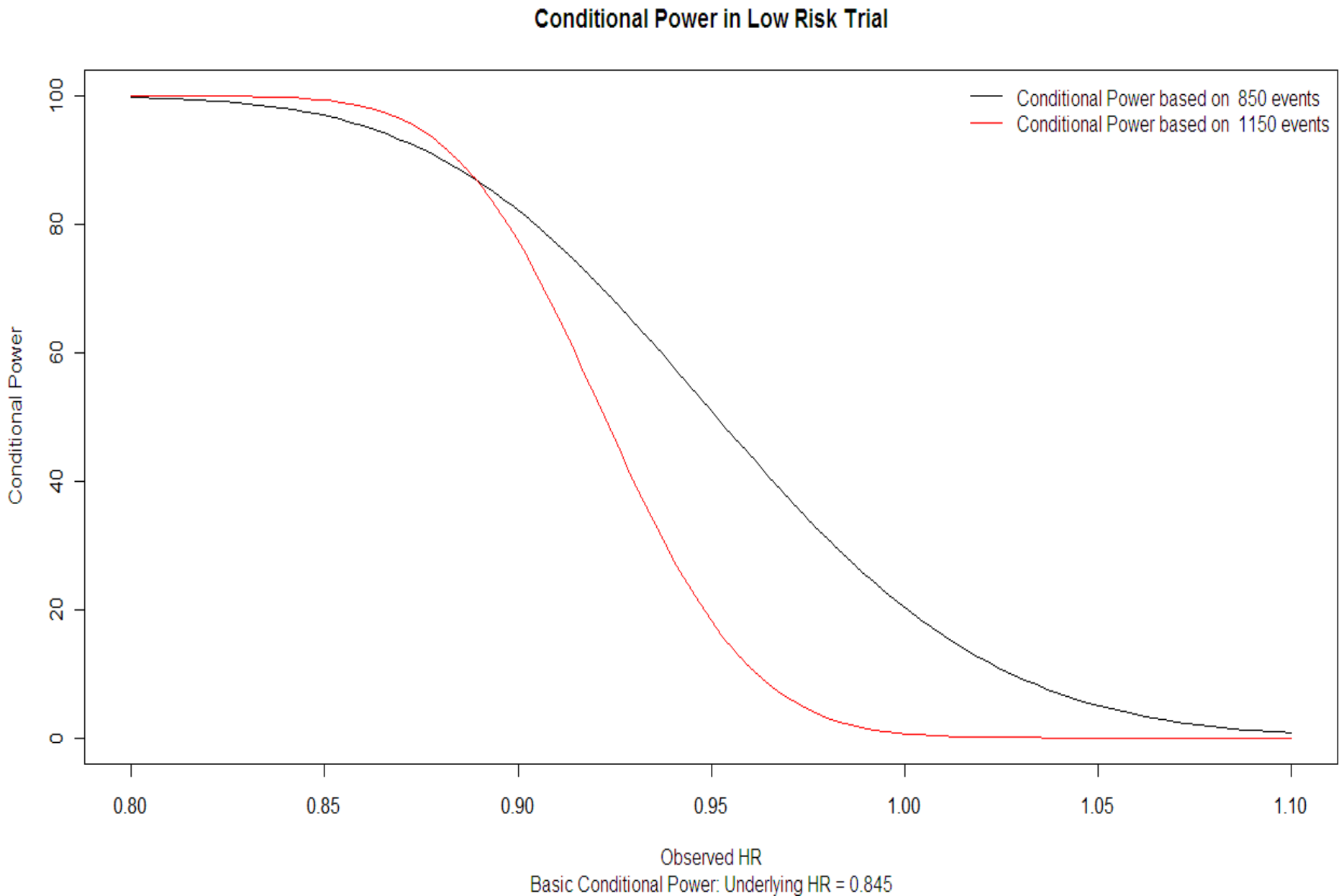
$\Sigma_{pred} = 4((m_0 + m)^{-1} + (m_{total})^{-1})$, *$m_{total}$ is the target number of events*

Modeling Log(HR) with Normal Distributions

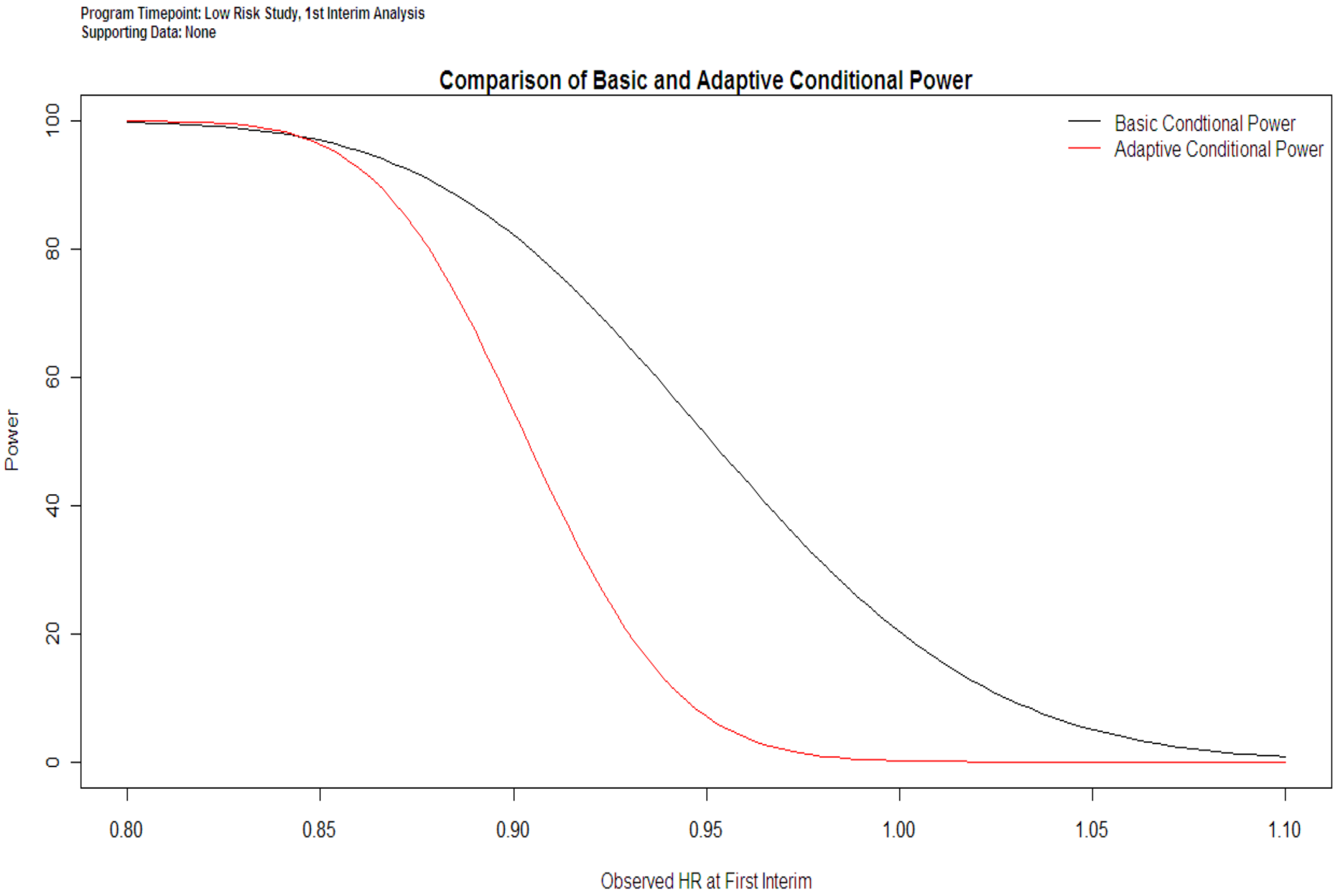
- Low Risk Trial
  - Recall this is the first trial
  - Use a non-informative prior to model the log(HR)
    - Take: $\theta_{prior,LRT} = 0$ (or log(0.845)) , $m_{0,LRT} = .0001$
    - Resulting predictive distribution is centered at the observed low risk log(HR).

- Let's contrast basic conditional power, adaptive conditional power and predictive power.

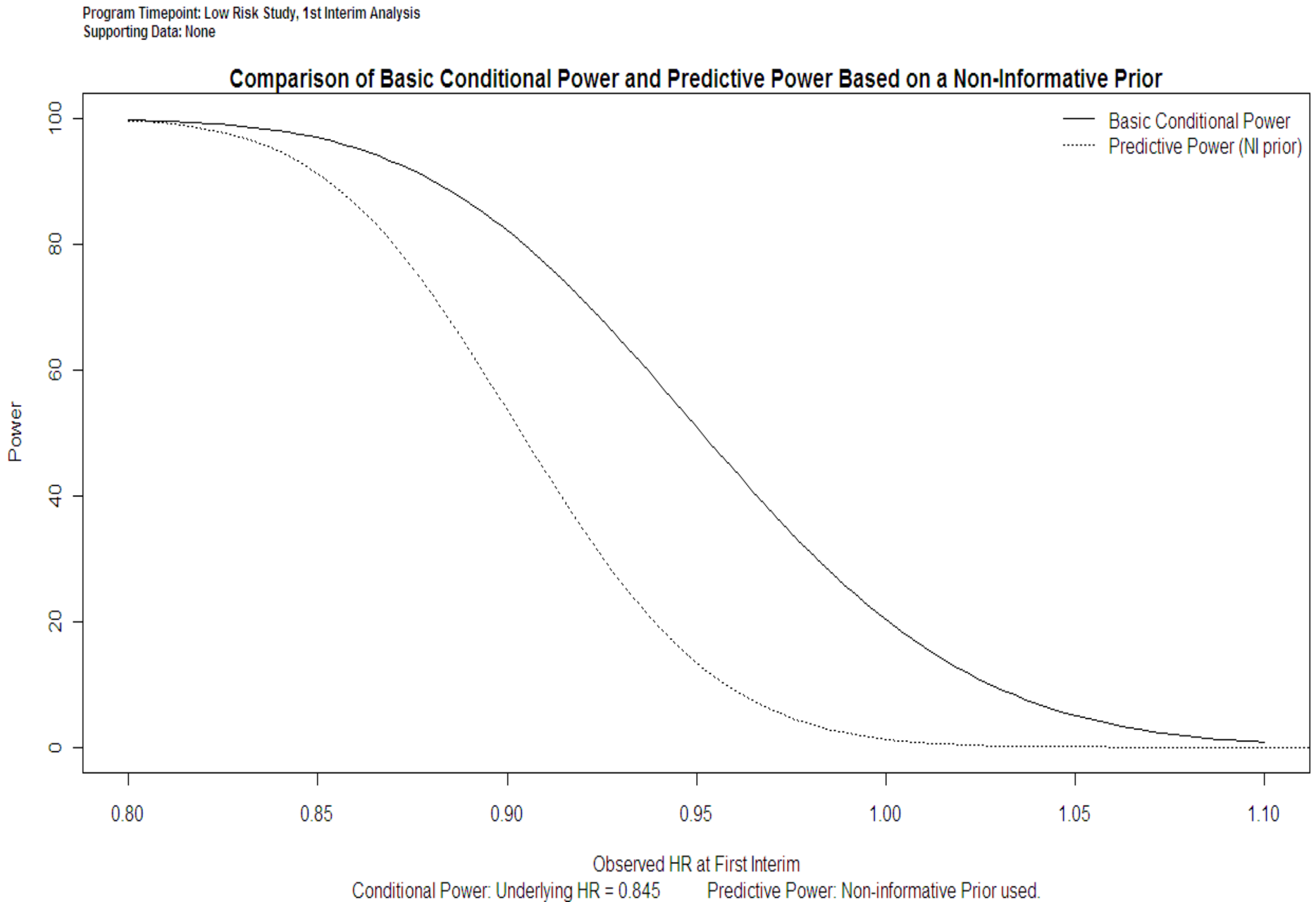
Comparison of Basic Conditional Power at the LRT IAs



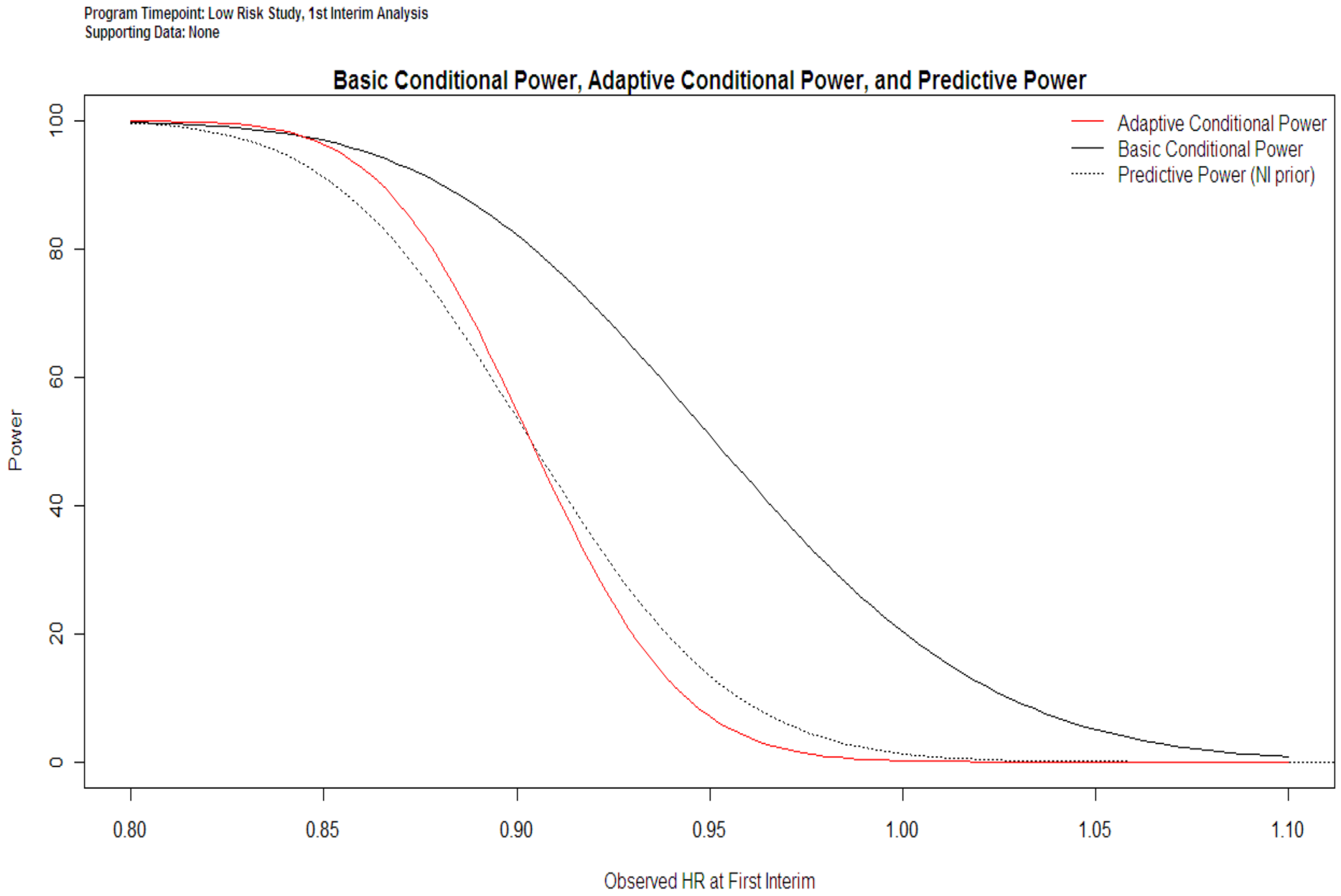
Conditional Power at 1st IA of LRT



Conditional Power and Predictive Power at 1st IA of LRT

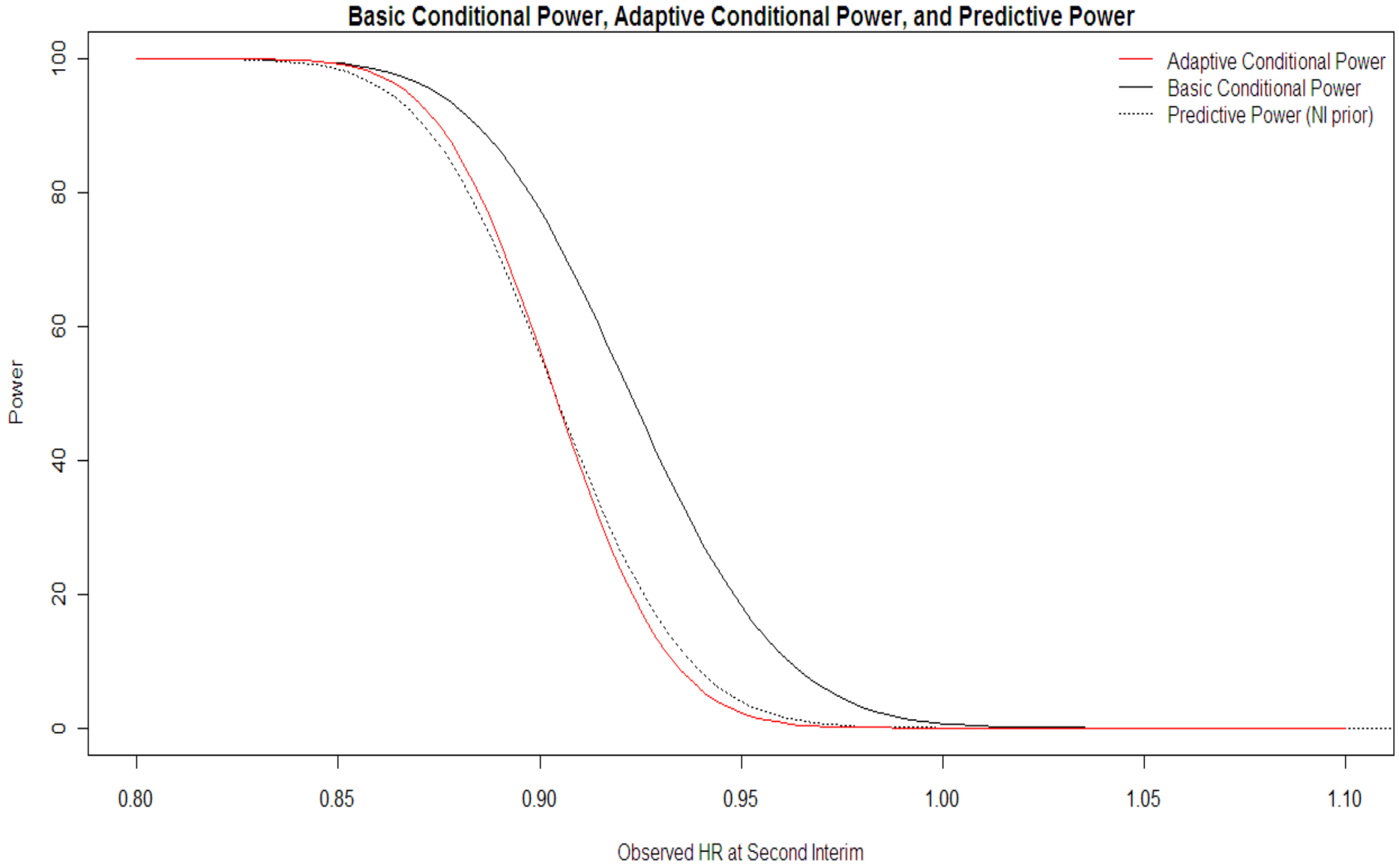


Different Types of Power at 1st Interim of LRT

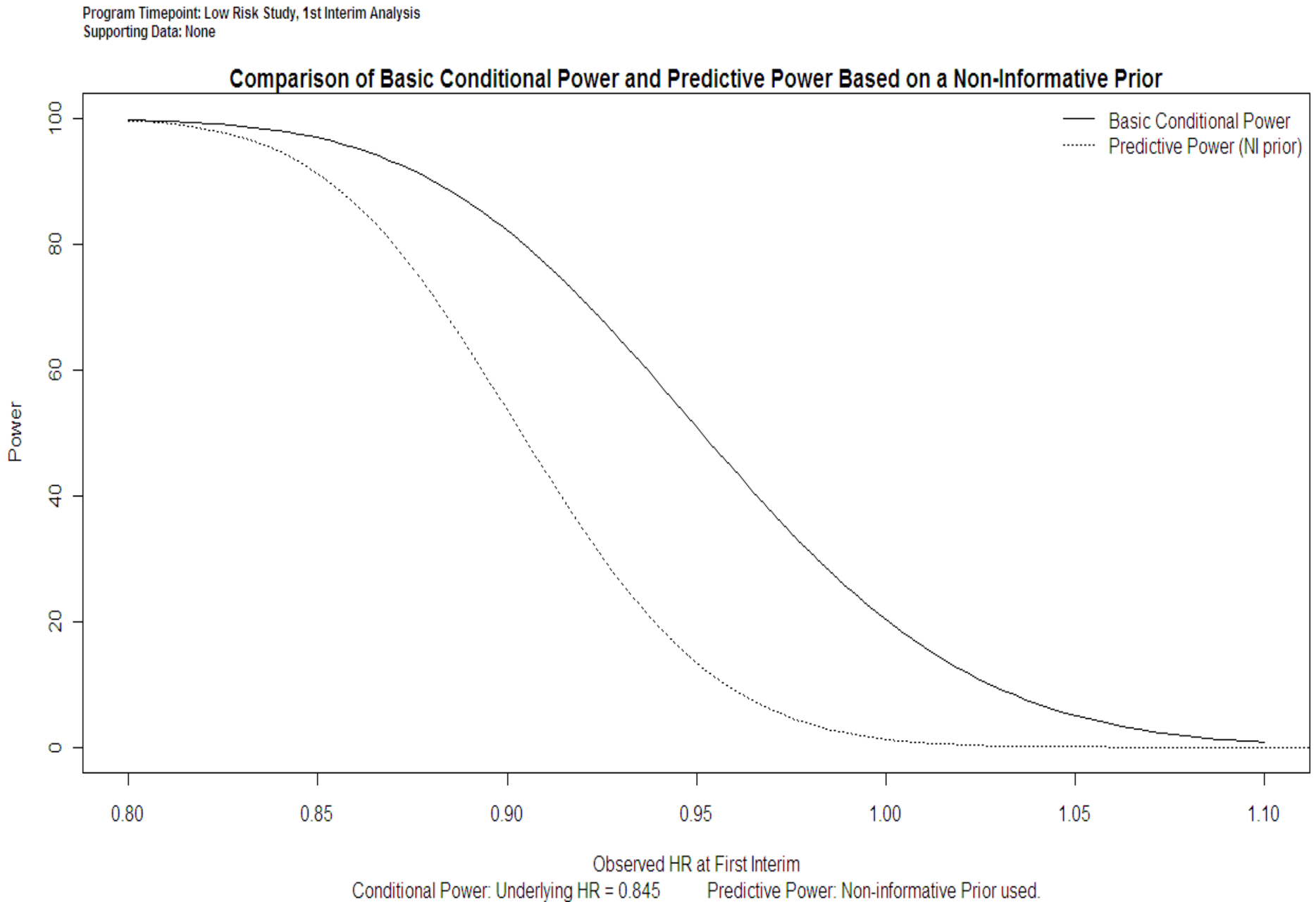


Different Types of Power at 2nd Interim of LRT

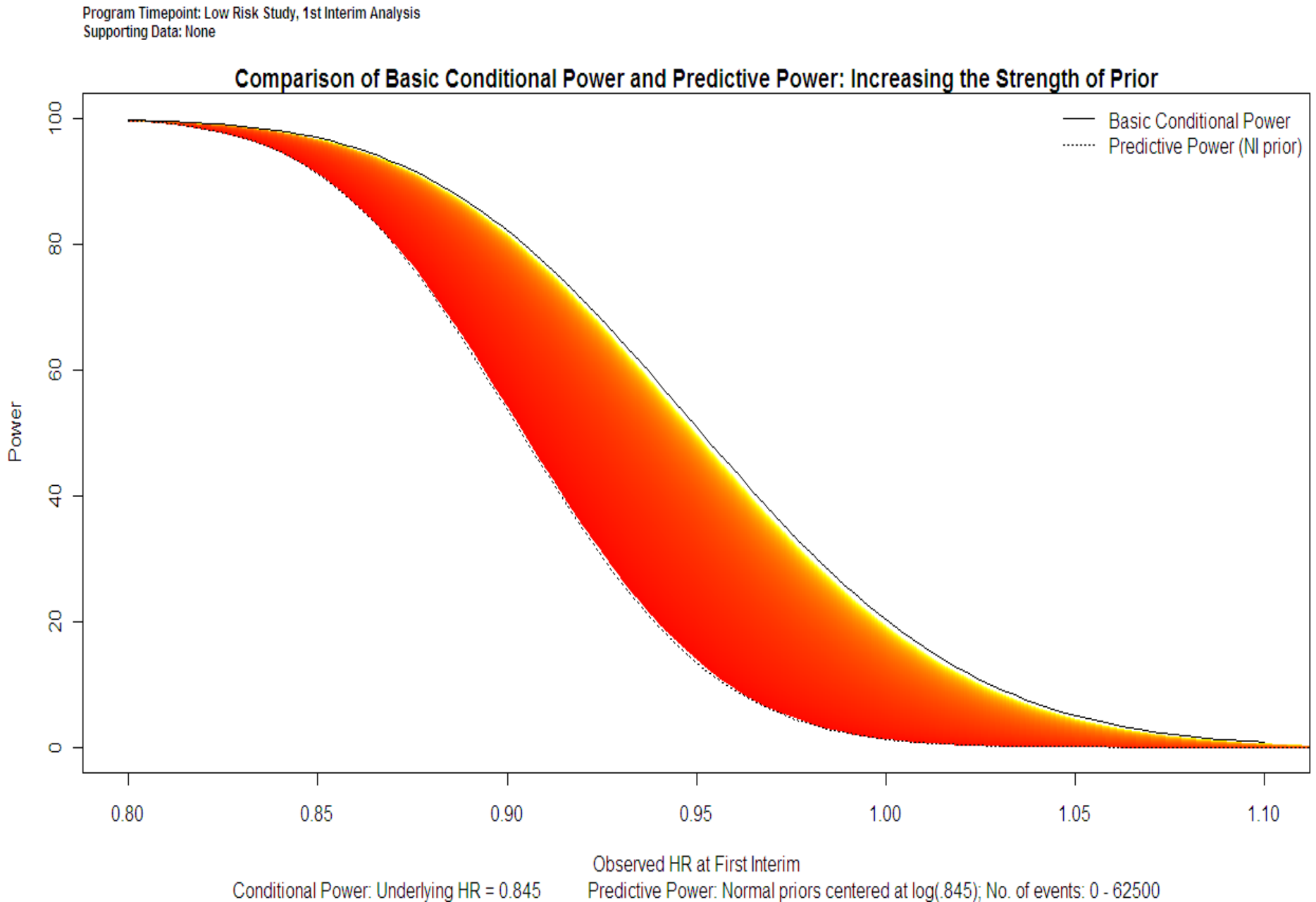
Program Timepoint: Low Risk Study, 2nd Interim Analysis
Supporting Data: None



Conditional Power and Predictive Power at 1st IA of LRT



Connecting Ideas: Predictive Power → Conditional Power



Harvest

- When observed HR is in the neighborhood of the alternative, power curves are in agreement
- When observed HR is greater than the alternative
  - Basic conditional power dominates adaptive conditional power and predictive power
  - Adaptive conditional power and predictive power have better agreement
- As information increases, power curves converge.
- By taking normal priors centered at the alternative, predictive power converges to basic conditional power as the number of ‘prior events’ increases.

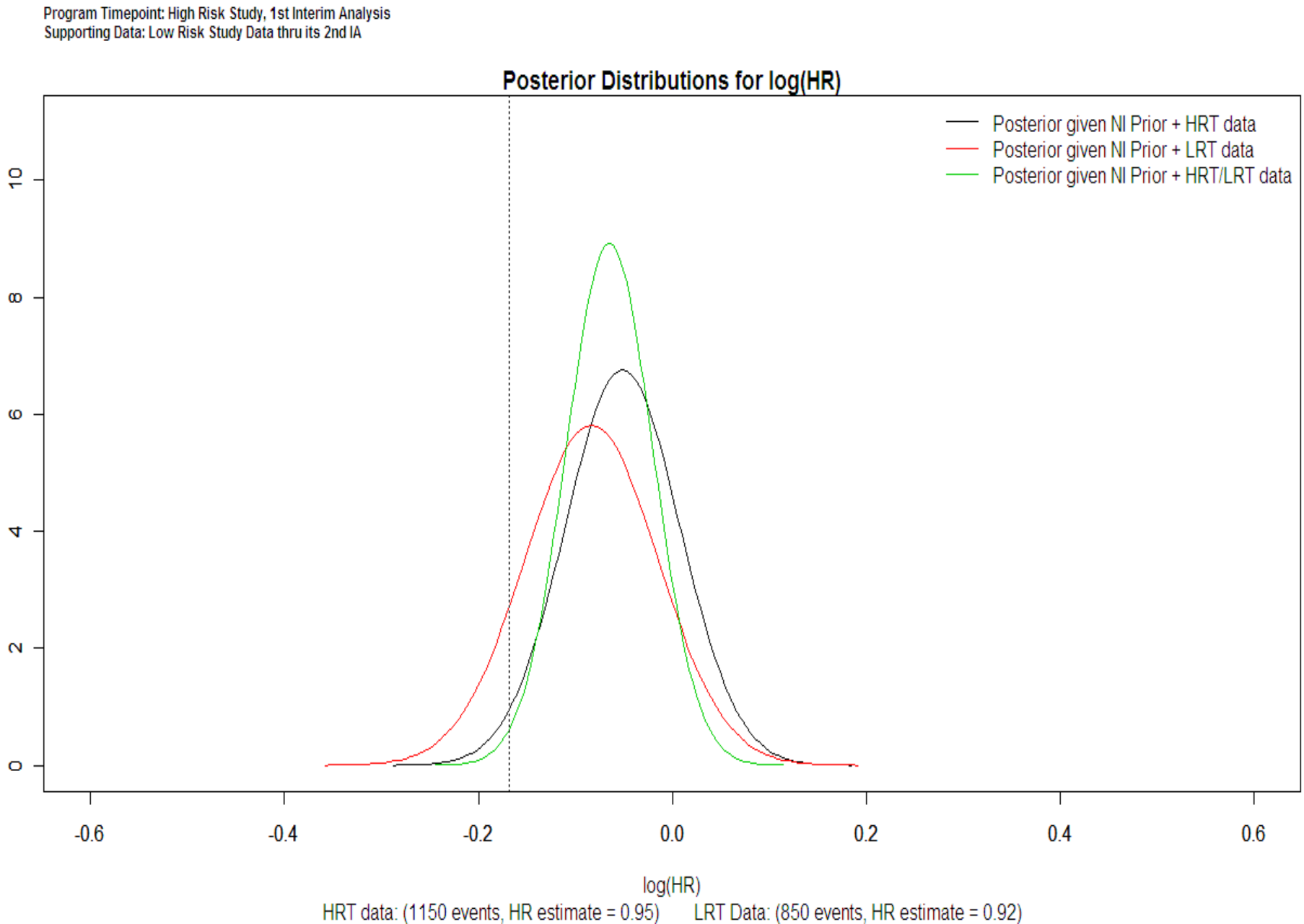
Modeling Log(HR) with Normal Distributions

- High Risk Trial
  - Since this trial has a delayed start, accumulated data from the LRT can be used.
    - Take $\theta_{prior,HRT} = \theta_{OBS,LRT}$, $m_{0,HRT} = m_{LOW}$
    - Resulting predictive distribution is centered at a point between the $\theta_{OBS,LRT}$ and $\theta_{OBS,HRT}$.

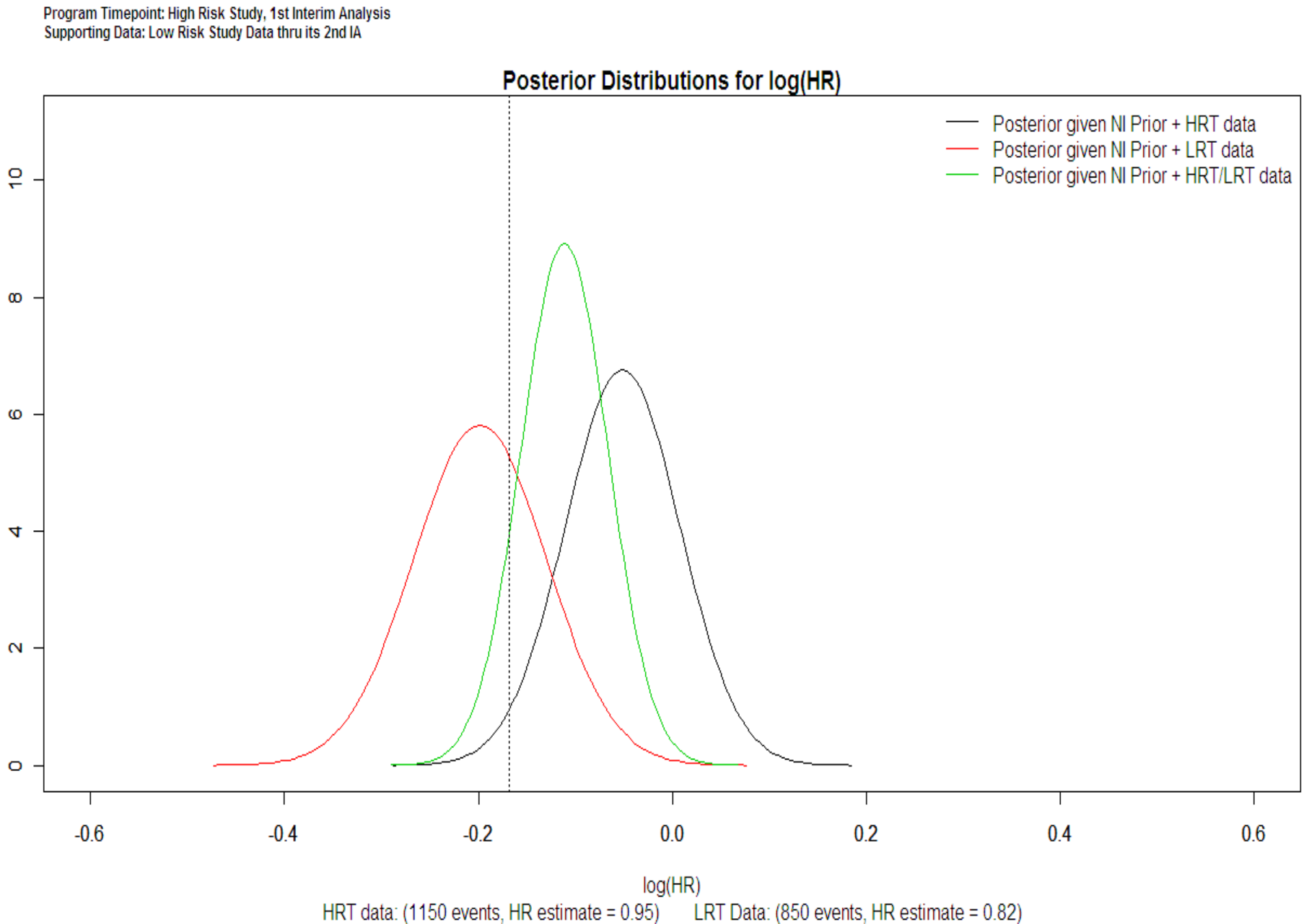
- We'll consider data available at HRT IAs
  - Program time point: 1st Interim Analysis of HRT
    - LRT provides data through it's 2nd Interim Analysis (HR estimate from 1150 events)
  - Program time point: 2nd Interim Analysis of HRT
    - LRT provides data through its final analysis (HR estimate from 1500 events)

- Is pooling warranted in the first place?

Pooling Studies Is Reasonable



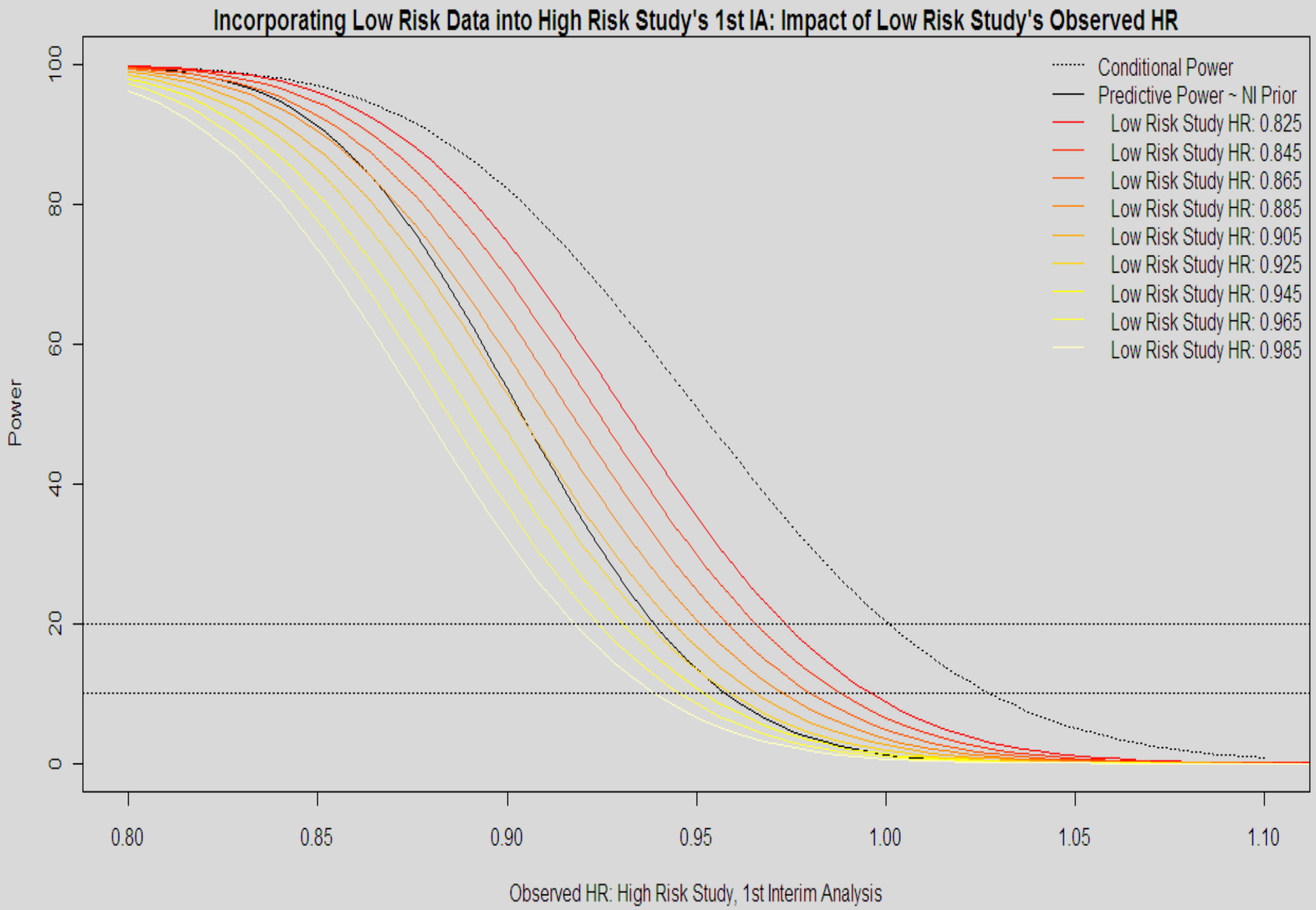
Pooling Is Not Reasonable



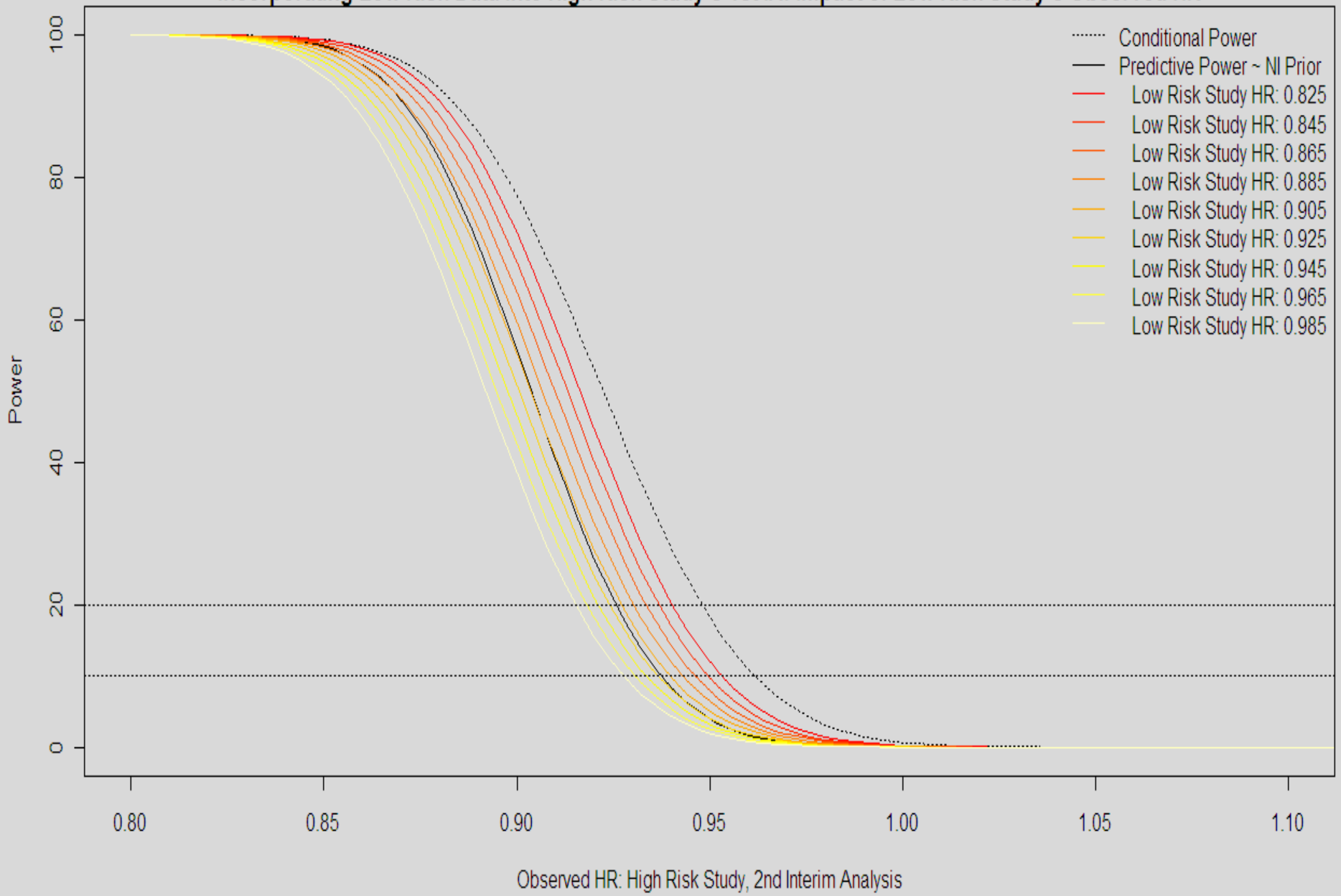
Proposal for Addressing Futility in HRT

- Provide IDMC with conditional power assessments as planned
- As in previous slides, review posterior distributions for treatment effects based on
  - LRT patients alone
  - HRT patients alone
  - LRT and HRT patients combined
- If posterior distributions **do not support** a common treatment effect IDMC can also review
  - Predictive power based on a non-informative prior
- If posterior distributions **do support** a common treatment effect IDMC can also review
  - Predictive power based on incorporating LRT data
  - Predictive power based on non-informative prior

Program Timepoint: High Risk Study, 1st Interim Analysis
Supporting Data: Low Risk Study Data thru its 2nd IA



Program Timepoint: High Risk Study, 2nd Interim Analysis
Supporting Data: Final Low Risk Study Data



Alternatives

- Begin with multiple priors
  - Suppose HR_{ALT} was the HR value used to power the trial
  - Skeptical Prior
    - Take $\theta_{prior} = 0$, variance taken so prior $P(HR < HR_{ALT}) = 0.05$
  - Enthusiastic prior
    - Take $\theta_{prior} = \log(HR_{ALT})$, variance taken so $P(HR > 1) = 0.05$

- Overkill for the IDMC?

Summary

- Connections between predictive power and conditional power can be made by considering a family of normal priors with decreasing variance
- Thresholds for power need to be adjusted depending on use of conditional/predictive power.
  - Predictive power values tend to be much more conservative

- Methods described offer a simple way to combine the information from two ongoing studies in order to make an more informed decision regarding futility.

References

- Snapinn, S., Chen, M. Jiang, Q., Koutsoukos, T. Assessment of futility in clinical trials. Pharmaceut. Statist. 2006; 5: 273–281

- Spiegelhalter, D., Abrams, K. Myles, J. 2004. Bayesian Approaches to Clinical Trials and Health-Care Evaluation
  - Speigelhalter, D., Freedman, L., Parmar, M. 1994. Bayesian Approaches to Randomized Trials. J. R. Statist. Soc. A. 157, 357-416
  - Parmar, M., Gareth, G., Speigelhalter, D., Shouhami, R., Altman, D. 2001. Monitoring of large randomized clinical trials: a new approach with Bayesian Methods. Lancet. 358: 375–81
  - Parmar, M., Speigelhalter, D., Freedman, L., 1994. The CHART Trials: Bayesian design and monitoring in practice. Statistics in Medicine. 13 1297-1312.
  - Tsiatis, A. A. (1981) The asymptotic joint distribution of the efficient scores test for the proportional hazards model calculated over time. Biometrika, 68, 311–15.