Interpretation of Safety Data in Clinical Trials

Milton Packer, M.D.

Department of Clinical Sciences
University of Texas Southwestern
Medical Center at Dallas

Evaluation of Efficacy	Evaluation of Safety

Evaluation of Efficacy	Evaluation of Safety
Limited number of analyses	
Detailed analytical approach prespecified at start of study	
Adequate statistical power	
Rigorous control of type I and II error	

Evaluation of Efficacy	Evaluation of Safety
Limited number of analyses	Extremely large number of analyses
Detailed analytical approach prespecified at start of study	Unlimited number of permutations and combinations
Adequate statistical power	Inadequate statistical power
Rigorous control of type I and II error	No meaningful control of type I and II error

Evaluation of Efficacy	Evaluation of Safety
Limited number of analyses	Extremely large number of analyses
Detailed analytical approach prespecified at start of study	Unlimited number of permutations and combinations
Adequate statistical power	Inadequate statistical power
Rigorous control of	No meaningful control of

Meaning of P value is well established

P value is difficult to interpret

Things to Worry About When Analyzing Incidence of Adverse Events

Things to Worry About When Analyzing Incidence of Adverse Events

1. There are hundreds of adverse events (multiplicity of comparisons).

Multiplicity of Comparisons

- A typical large-scale clinical trial may describe as many as 500 individual terms describing adverse events.
- If P value were calculated for each pairwise comparison, then one would — by chance alone — expect

$\Rightarrow$ 25 events (5%) to have $P \leq 0.05$

$\Rightarrow$ 5 events (1%) to have $P \leq 0.01$

Things to Worry About When Analyzing Incidence of Adverse Events

1. There are hundreds of adverse events (multiplicity of comparisons).
2. Adverse events are spontaneous (nonadjudicated) reports.

Adverse Events Are Spontaneous (Nonadjudicated) Reports

- Adverse events are reported at the discretion of the investigator and then translated into standardized terms.
- There is little uniformity as to how an event is identified, defined or reported.
- Uncertainty increased when event is in field remote from investigator's focus.

Can We Fix This Problem by Blinded Post Hoc Adjudication?

Can We Fix This Problem by Blinded Post Hoc Adjudication?

- Rules guiding post hoc adjudication are inevitably influenced by knowledge that a treatment effect has been seen.
  - Any bar set by the post hoc process can magnify or dilute the effect.
- Adjudication is generally not applied to confirm absence of event.

Things to Worry About When Analyzing Incidence of Adverse Events

1. There are hundreds of adverse events (multiplicity of comparisons).
2. Adverse events are spontaneous (nonadjudicated) reports.
3. Analyses that depend on grouping of events are subject to bias.

Analyses That Depend on Grouping of Adverse Events

- Those creating the grouping have frequently already looked at the data and know (subconsciously) what kind of definition is needed to capture the events of interest.

Thrombotic Cardiovascular Events

	Placebo	Drug
Myocardial infarction	5	20
Stroke	4	4
Sudden death	5	6
Unstable angina	6	5
Pulmonary embolism	3	4
Arterial embolism	0	1
Transient ischemic attack	5	3
Venous thrombosis	4	6

Thrombotic Cardiovascular Events

	Placebo	Drug
Myocardial infarction	5	10
Stroke	4	8
Sudden death	5	6
Unstable angina	6	5
Pulmonary embolism	3	4
Arterial embolism	0	1
Transient ischemic attack	5	3
Venous thrombosis	4	6

Thrombotic Cardiovascular Events

	Placebo	Drug
Myocardial infarction	5	4
Stroke	4	4
Sudden death	5	9
Unstable angina	6	13
Pulmonary embolism	3	5
Arterial embolism	0	1
Transient ischemic attack	5	11
Venous thrombosis	4	6

Thrombotic Cardiovascular Events

	Placebo	Drug
Myocardial infarction	5	4
Stroke	4	4
Sudden death	5	9
Unstable angina	6	13
Pulmonary embolism	3	5
Arterial embolism	0	1
Transient ischemic attack	5	11
Venous thrombosis	4	6

Analyses That Depend on Grouping of Adverse Events

- Best to develop uniform definition of a “group” before classifying events.
- When the process of developing a definition is started after a concern has been raised, those creating the definition have frequently already looked at the data and know (subconsciously) what kind of definition is needed to capture the events of interest.

Things to Worry About When Analyzing Incidence of Adverse Events

1. There are hundreds of adverse events (multiplicity of comparisons).
2. Adverse events are spontaneous (nonadjudicated) reports.
3. Analyses that depend on grouping of events are subject to bias.
4. Small number of events results in extremely imprecise estimates.

With Small Number of Events, Lack of an Observed Difference Does *Not* Rule Out Existence of True Difference

Major Adverse Cardiovascular Event

Placebo	Drug
25	26

RR=1.04
95% CI (0.60-1.79)
P = 0.89

With Small Number of Events, the Finding of an Observed Difference Does *Not* Prove Existence of True Difference

Major Adverse Cardiovascular Event

Placebo	Drug
13	33

RR=2.63
95% CI (1.39, 5.00)
P = 0.002

What Is Wrong With Imprecise Estimates?

What Is Wrong With Imprecise Estimates?

- Imprecise estimates are fine if the intent is to withhold judgment until more data are collected to make the estimates more precise.

What Is Wrong With Imprecise Estimates?

- Imprecise estimates are fine if the intent is to withhold judgment until more data are collected to make the estimates more precise.
- Imprecise estimates are problematic if the intent is to stop and reach a conclusion.

What Is Wrong With Imprecise Estimates?

- Imprecise estimates are fine if the intent is to withhold judgment until more data are collected to make the estimates more precise.
- Imprecise estimates are problematic if the intent is to stop and reach a conclusion.
- When calculated in the conventional manner, the 95%CIs (and the associated P value) of an estimate have meaning primarily in the context of a completed experiment.

What Is Wrong With Imprecise Estimates?

- The adverse event data generated in a typical trial is not the result of a completed experiment.

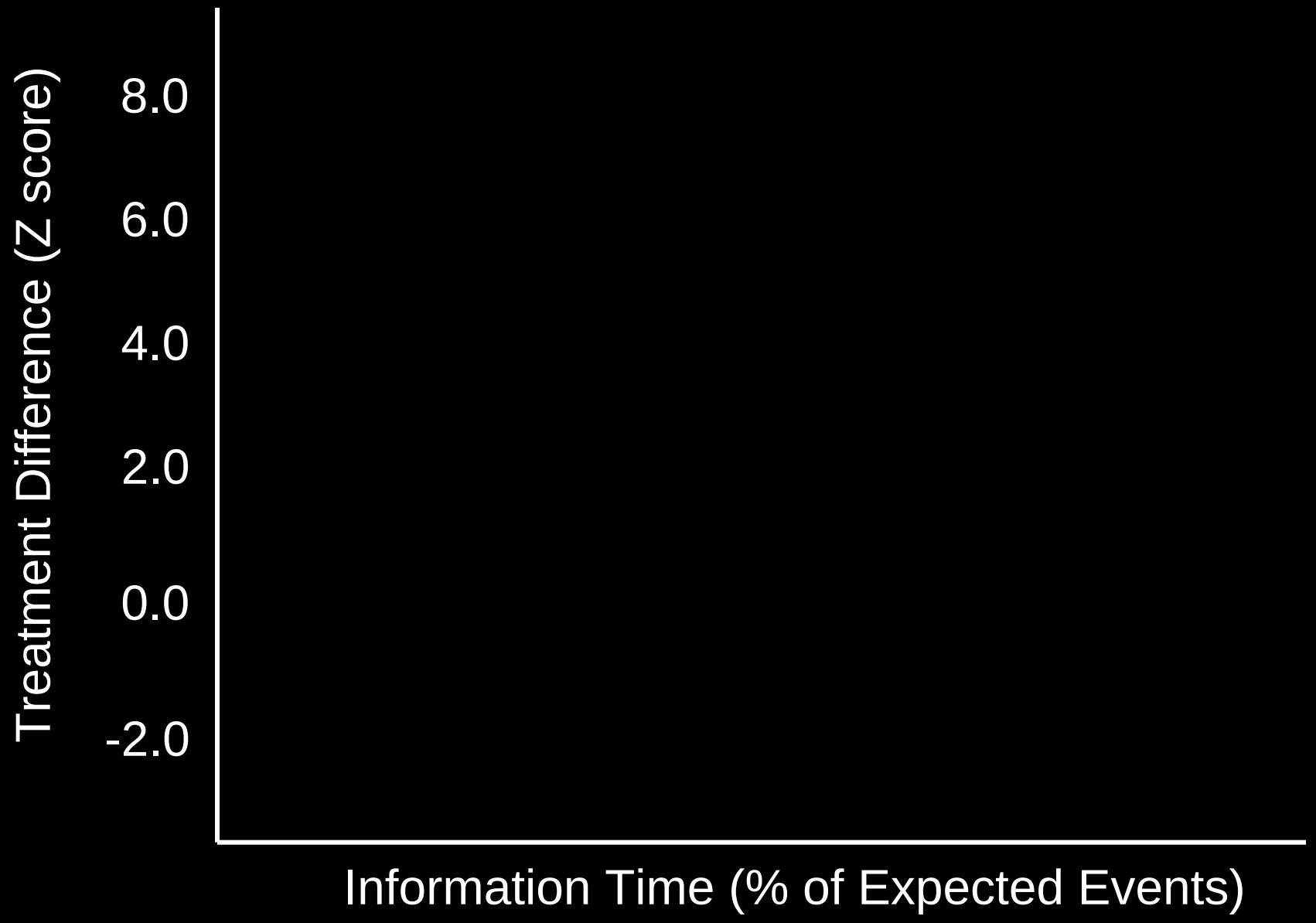
What Is Wrong With Imprecise Estimates?

- The adverse event data generated in a typical trial is not the result of a completed experiment.
- Viewed from the amount of data needed for a precise estimate, the adverse event data in a single study represents a snapshot in an ongoing experiment to characterize the safety of the drug.

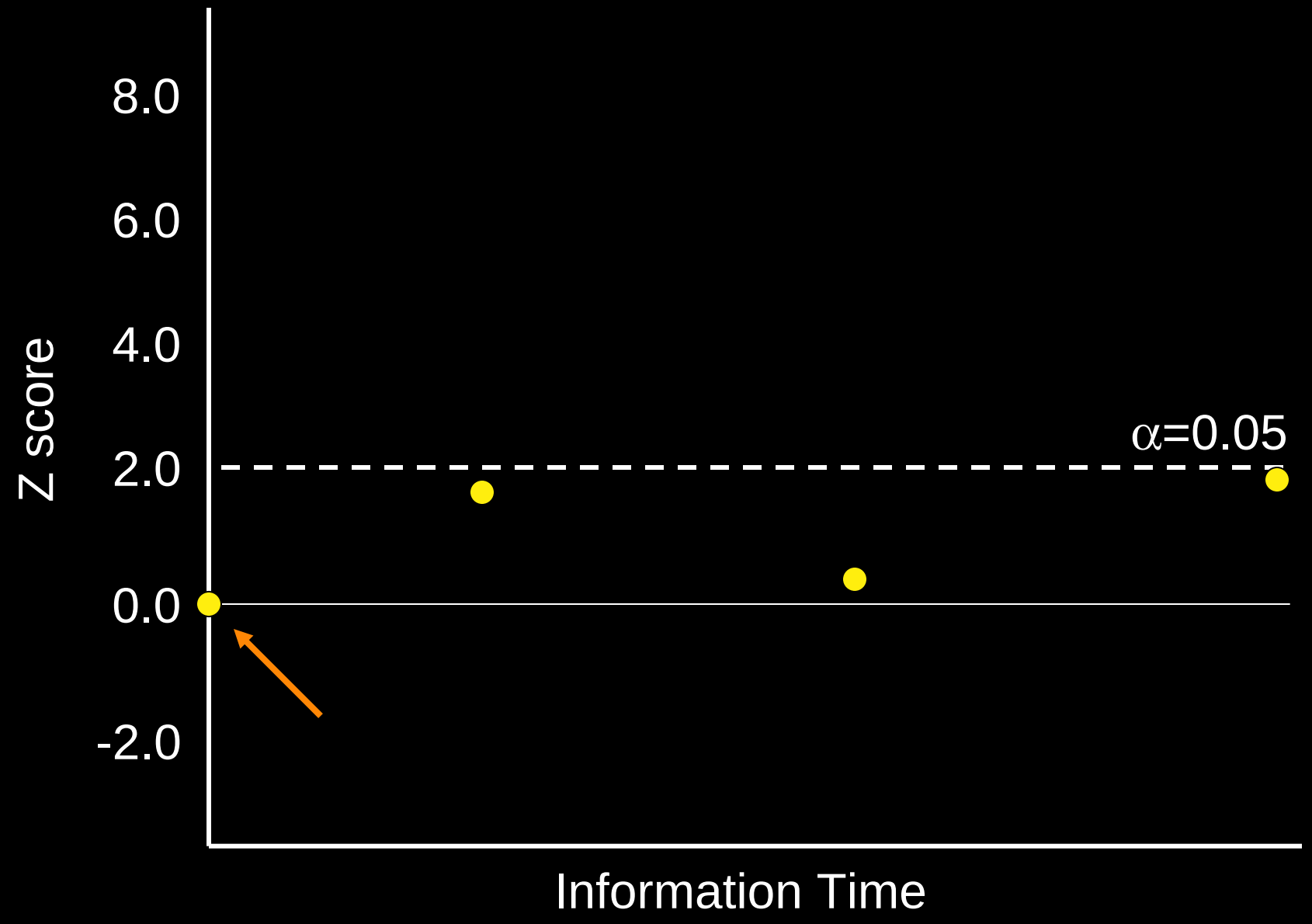
What Is Wrong With Imprecise Estimates?

- The adverse event data generated in a typical trial is not the result of a completed experiment.
- Viewed from the amount of data needed for a precise estimate, the adverse event data in a single study represents a snapshot in an ongoing experiment to characterize the safety of the drug.
- Therefore, performing an analysis of adverse events data is akin to interim analyses of primary endpoint data in an ongoing clinical trial.

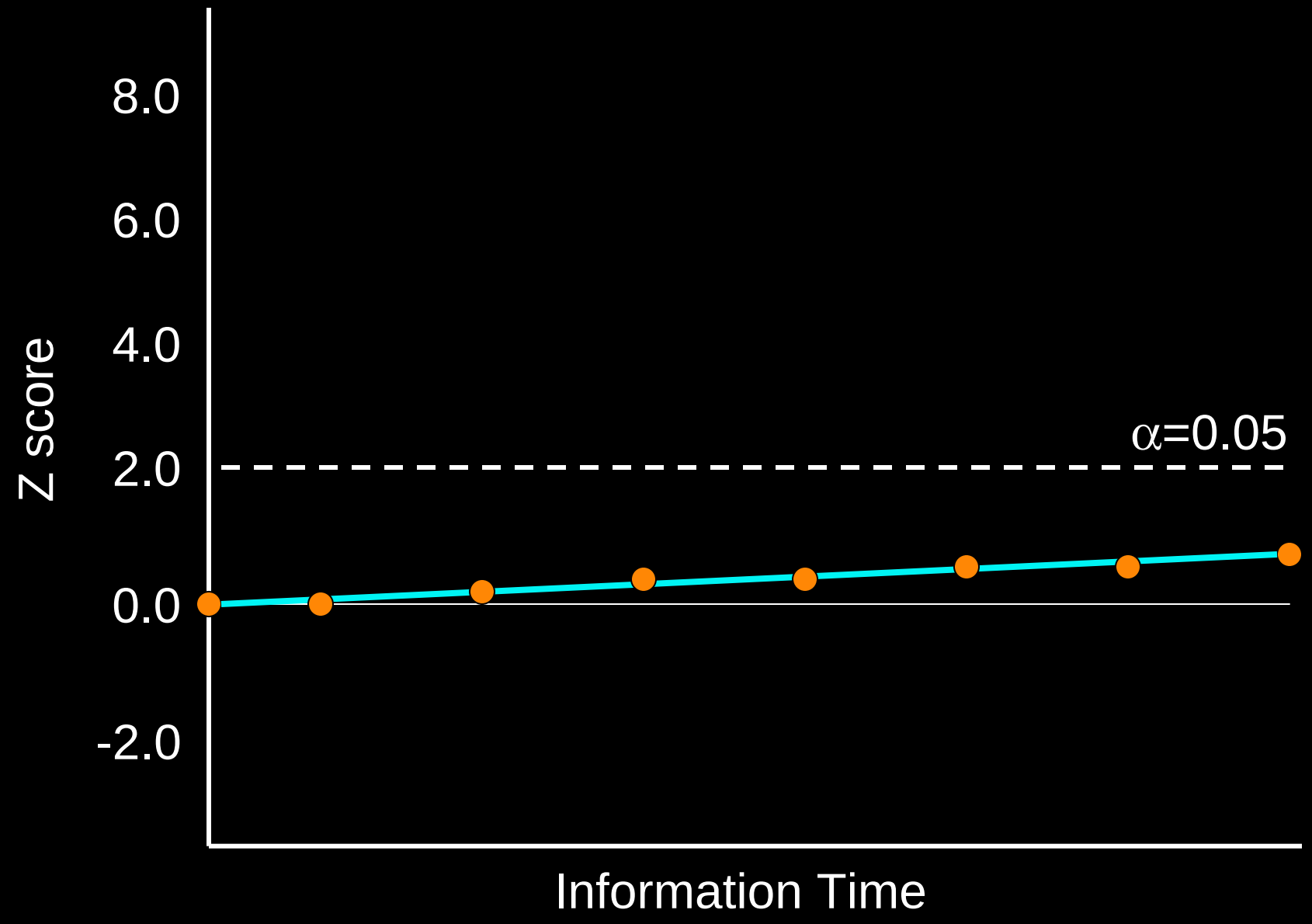
Interim Monitoring in Group Sequential Trials



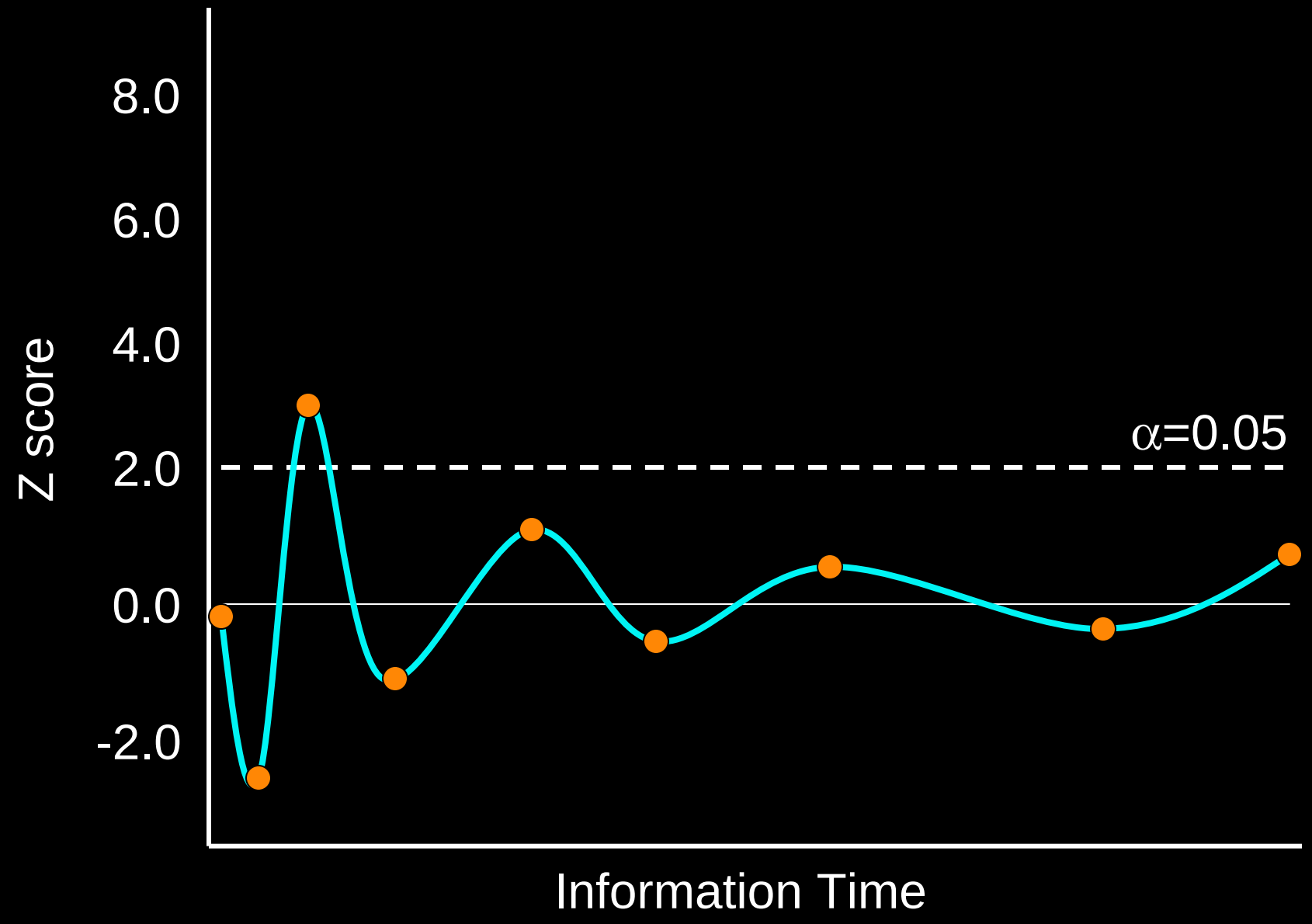
Interim Monitoring in Group Sequential Trials



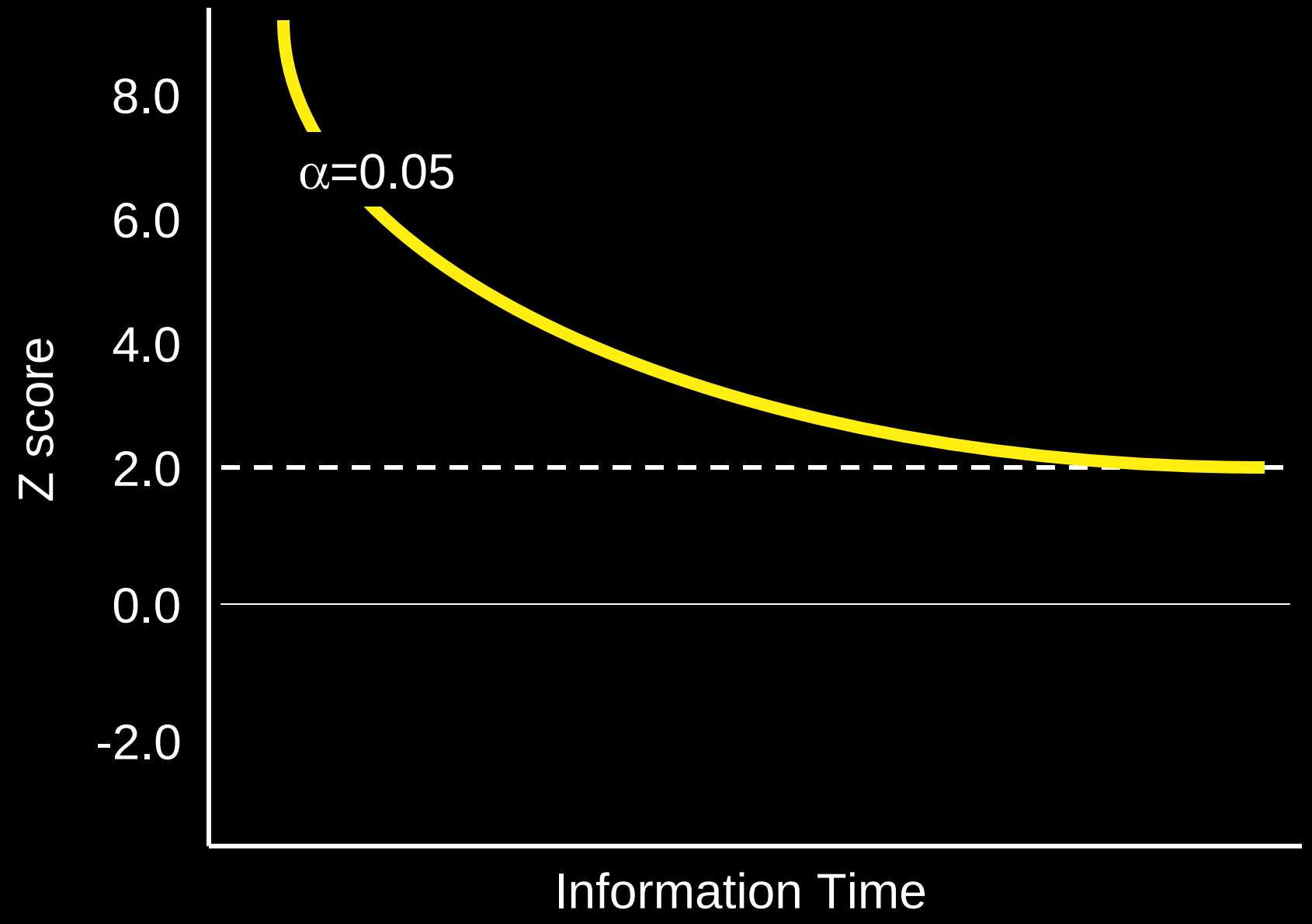
Interim Monitoring in Group Sequential Trials



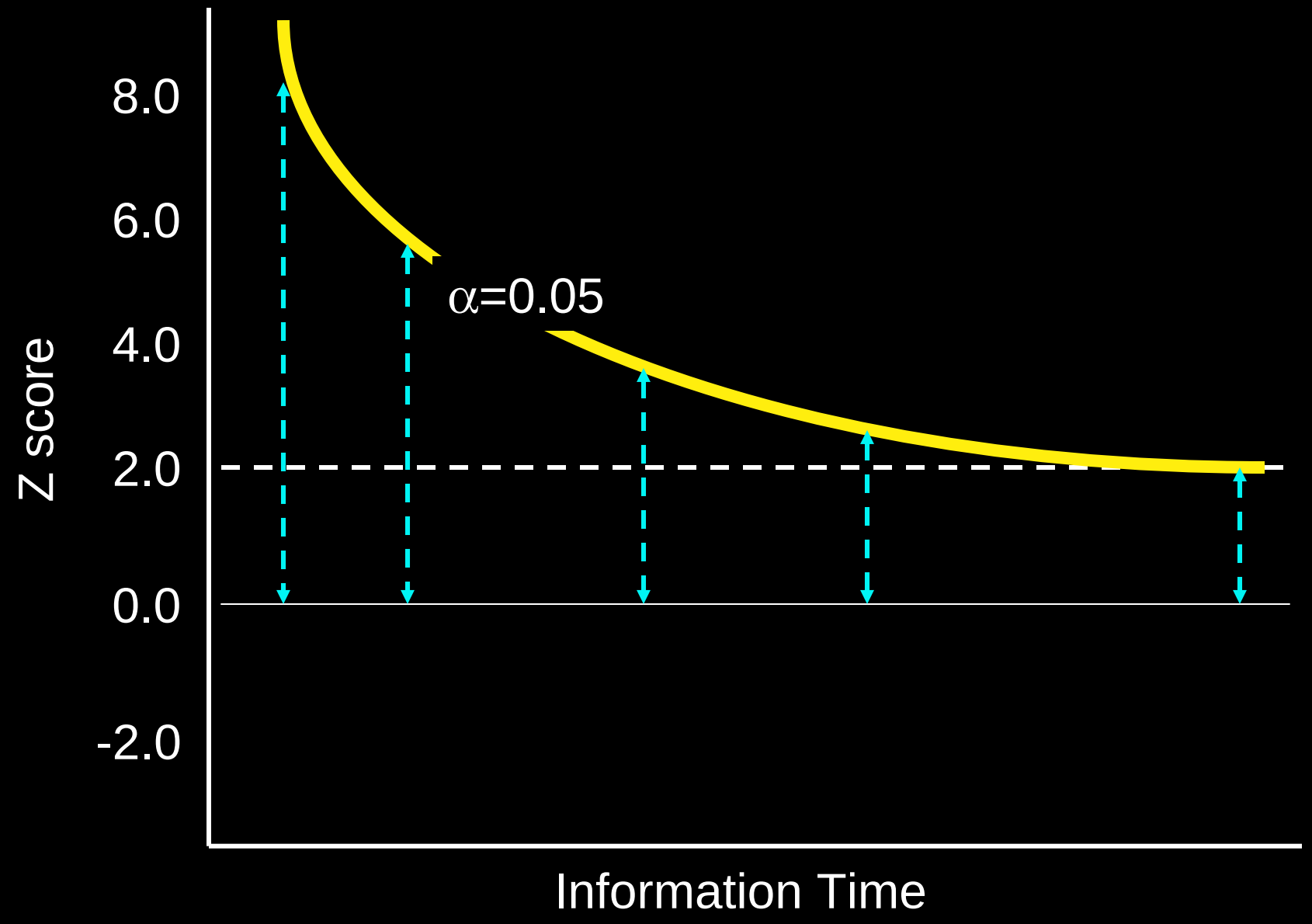
Interim Monitoring in Group Sequential Trials



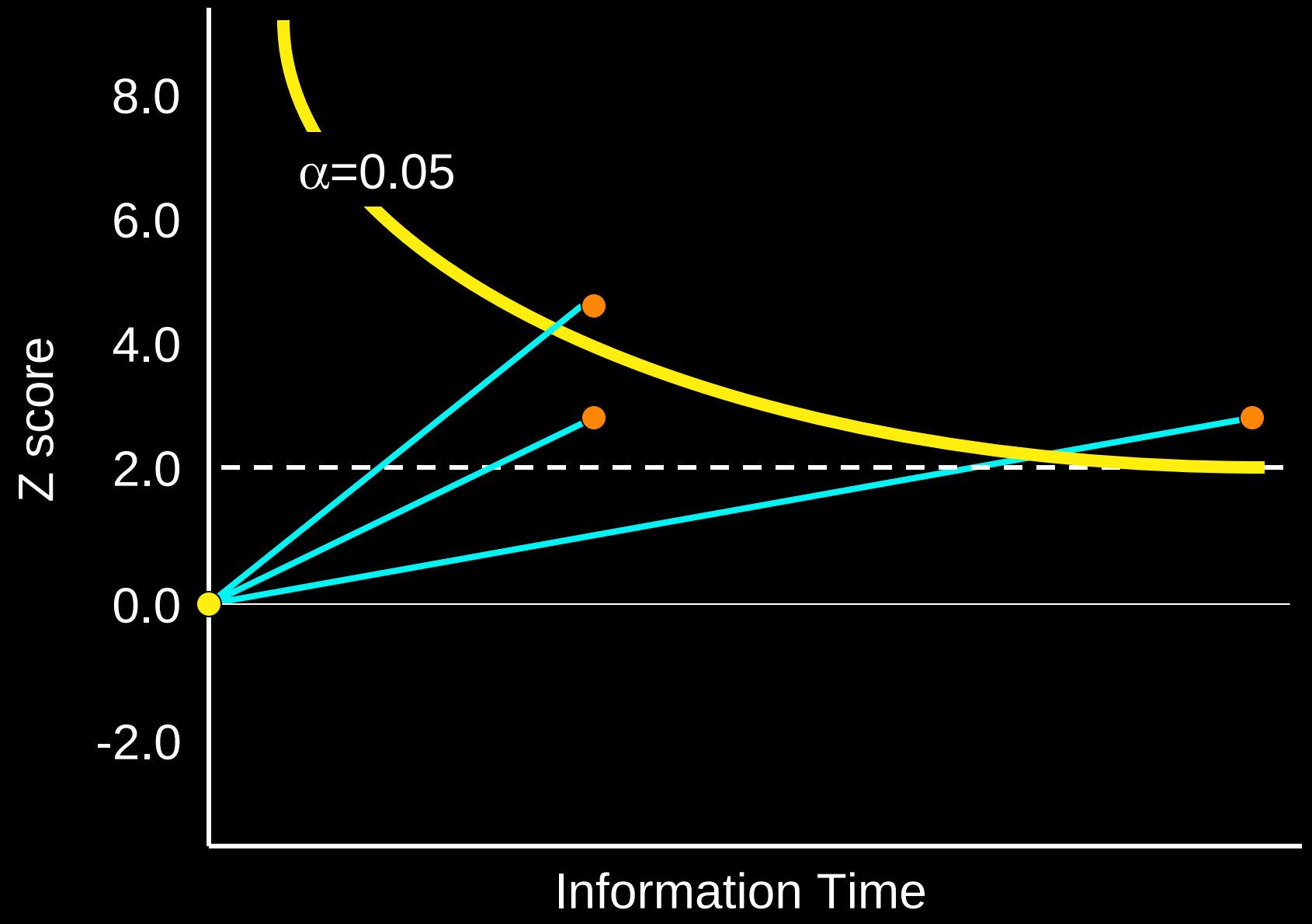
Interim Monitoring in Group Sequential Trials



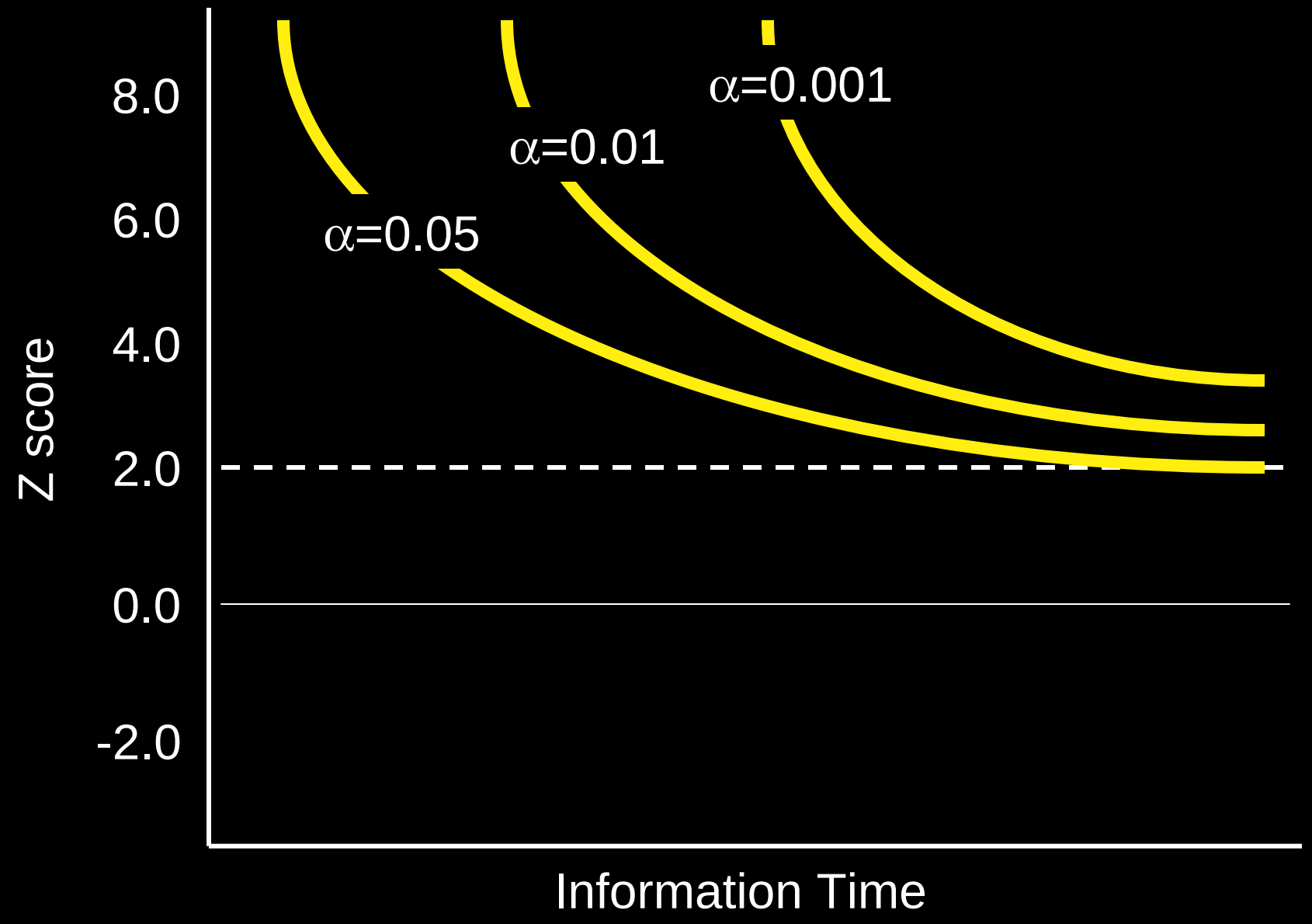
Interim Monitoring in Group Sequential Trials



Interim Monitoring in Group Sequential Trials



Interim Monitoring in Group Sequential Trials

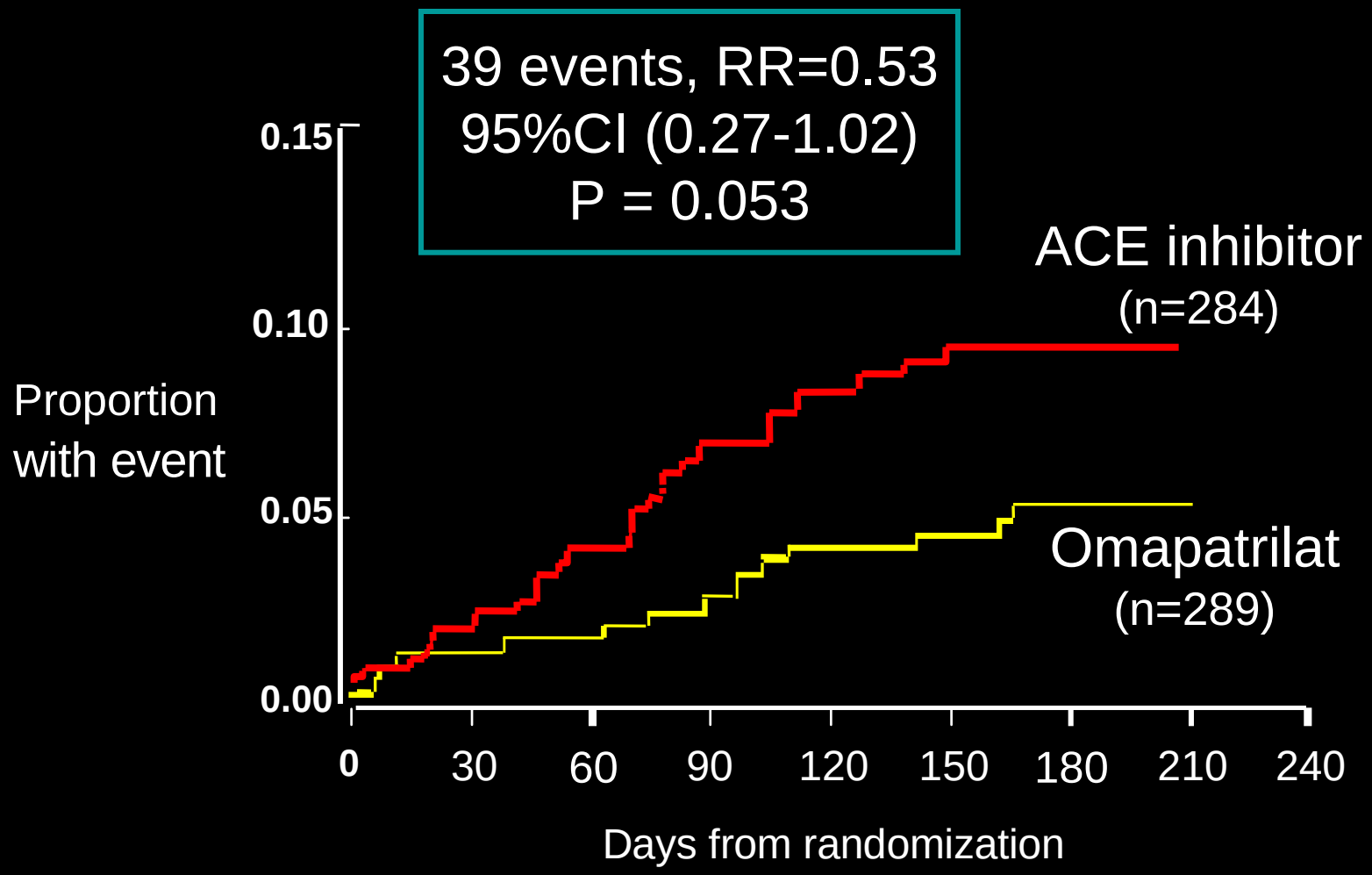


Concept

- Reaching conclusions from data derived in an underpowered trial raises the same concerns as reaching conclusions based on an underpowered interim analysis in a definitive, adequately powered trial.

IMPRESS (Omapatrilat in Heart Failure)

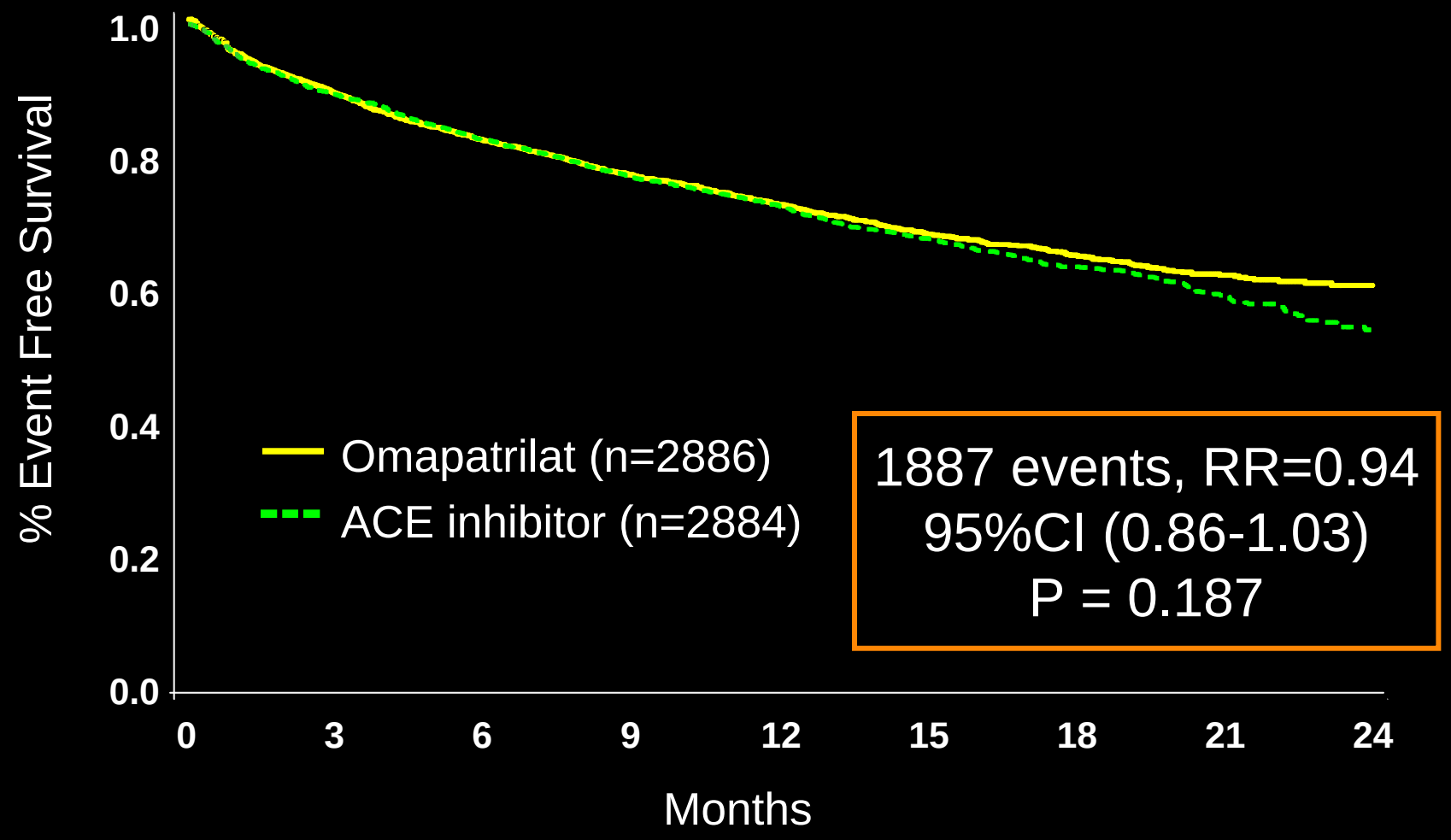
Death or Hospitalization for Heart Failure



Rouleau et al. Lancet 2000; 356:615-20.

OVERTURE Trial

Death or Hospitalization for Heart Failure



Packer et al. Circulation 2002: 106:920-6.

Amlodipine in Heart Failure

All-Cause Mortality in Nonischemic Cardiomyopathy

	Placebo	Amlodipine	Hazard Ratio	Log-rank P-Value
PRAISE-1	74/212	45/209	0.55 (0.37,0.79)	0.001

Amlodipine in Heart Failure

All-Cause Mortality in Nonischemic Cardiomyopathy

	Placebo	Amlodipine	Hazard Ratio	Log-rank P-Value
PRAISE-1	74/212	45/209	0.55 (0.37,0.79)	0.001
PRAISE-2	262/826	278/826	1.09 (0.92, 1.29)	0.32

Experience With Vesnarinone and Losartan

		Control	Drug	Hazard ratio (95% CI)	P value
Vesnarinone (vs placebo)	Initial study	33/238	13/239	0.38 (0.20-0.72)	0.002
Losartan (vs captopril)	Initial study	32/370	17/352	0.54 (0.31-0.95)	0.035

N Engl J Med 1993;329:149-55. N Engl J Med 1998;339:1810-6.
Lancet 1997; 349: 747-57. Lancet 2000; 355:582-7.

Definitive Trial Shows Reversal of Effect Seen in Earlier Pilot Trial

		Control	Drug	Hazard ratio (95% CI)	P value
Vesnarinone (vs placebo)	Initial study	33/238	13/239	0.38 (0.20-0.72)	0.002
	Definitive study	242/1283	292/1275	1.22 (1.04-1.42)	0.02
Losartan (vs captopril)	Initial study	32/370	17/352	0.54 (0.31-0.95)	0.035
	Definitive study	250/1574	280/1578	1.13 (0.95-1.35)	0.16

N Engl J Med 1993;329:149-55. N Engl J Med 1998;339:1810-6.
Lancet 1997; 349: 747-57. Lancet 2000; 355:582-7.

Magnesium in Myocardial Infarction

All-Cause Mortality

	Placebo	Magnesium	Hazard Ratio	Log-rank P-Value
Meta-analysis	53/644	25/644	0.45 (0.28,0.71)	< 0.001
LIMIT-2	90/1159	118/1157	0.74 (0.55,1.00)	0.04

Magnesium in Myocardial Infarction

All-Cause Mortality

	Placebo	Magnesium	Hazard Ratio	Log-rank P-Value
Meta-analysis	53/644	25/644	0.45 (0.28,0.71)	< 0.001
LIMIT-2	90/1159	118/1157	0.74 (0.55,1.00)	0.04
ISIS-4	2103/29039	2216/29011	1.06 (1.00,1.12)	0.07

Metoprolol XL in Heart Failure

Effect on CHF Hospitalizations

	Placebo	Metoprolol	Hazard Ratio	Log-rank P-Value
RESOLVD	5/212	15/214	------ (1.01,5.63)	< 0.05

Metoprolol XL in Heart Failure

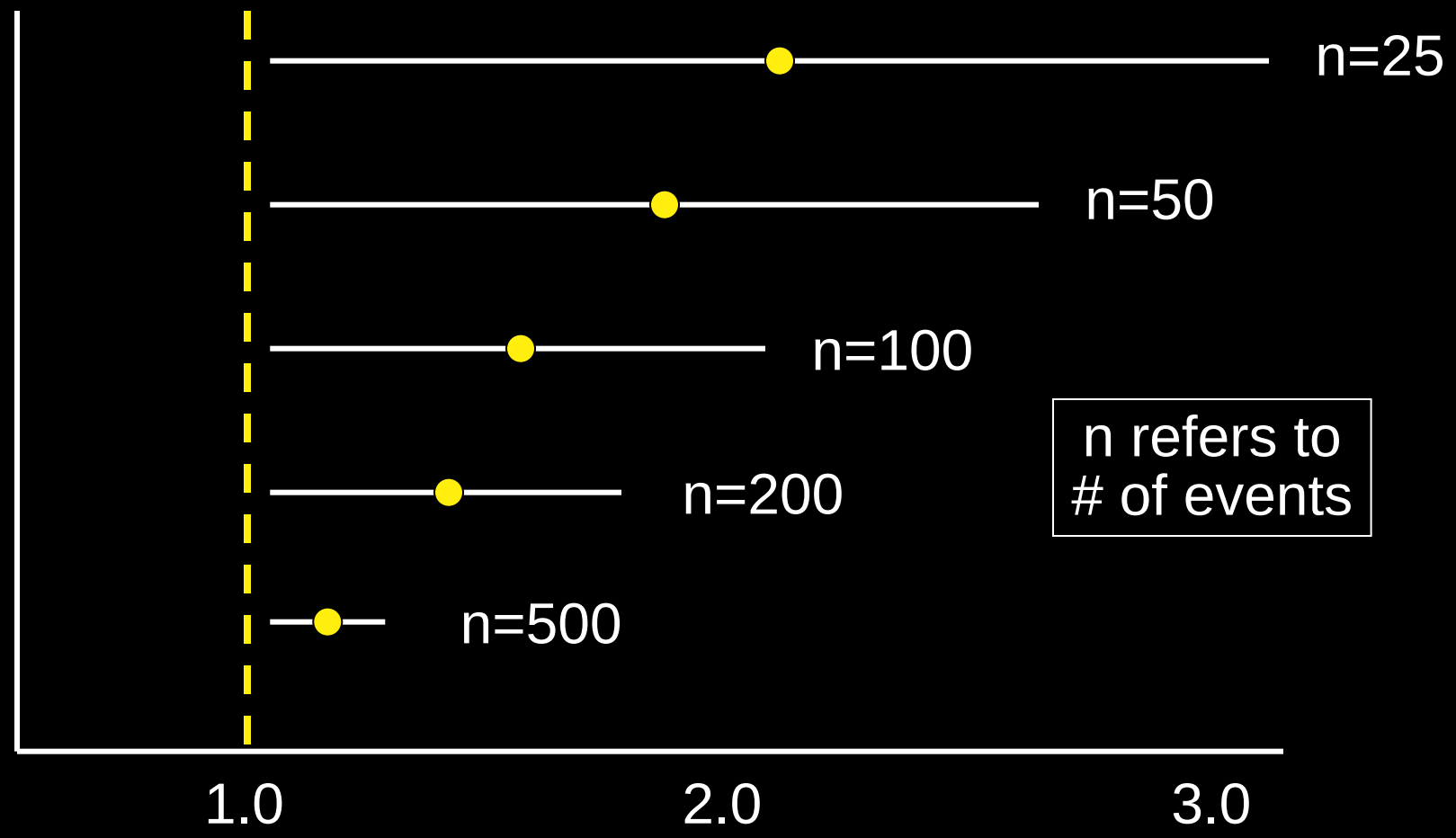
Effect on CHF Hospitalizations

	Placebo	Metoprolol	Hazard Ratio	Log-rank P-Value
RESOLVD	5/212	15/214	------ (1.01,5.63)	< 0.05
MERIT-HF	294/826	200/826	----- (-----, -----)	< 0.001

What We Have Learned

- To achieve statistical significance in an underpowered analysis, the effect size must be extreme and the estimate must be imprecise.

Effect Size and Confidence Intervals For Trials With Varying Number of Events Assuming Lower Bound ≈ 1.00 and Thus P ≤ 0.05



What We Have Learned

- To achieve statistical significance in an underpowered analysis, the effect size must be extreme and the estimate must be imprecise.
- Yet, the more extreme the effects and the more imprecise the estimates, the less likely the result will be reproduced in definitive clinical trials.

What To Do?

What To Do?

- The most important first step is to develop an approach to analyzing data in trials with small numbers of events which accurately reflects the true imprecision of the treatment effect estimate and its statistical significance.

How Should Such a Difference Between Treatment Groups Be Interpreted?

Major Adverse Cardiovascular Event

Placebo	Drug
13	33

How Should Such a Difference Between Treatment Groups Be Interpreted?

Major Adverse Cardiovascular Event

Placebo	Drug
13	33

RR=2.63
95% CI (1.39, 5.00)
P = 0.002

How Should Such a Difference Between Treatment Groups Be Interpreted?

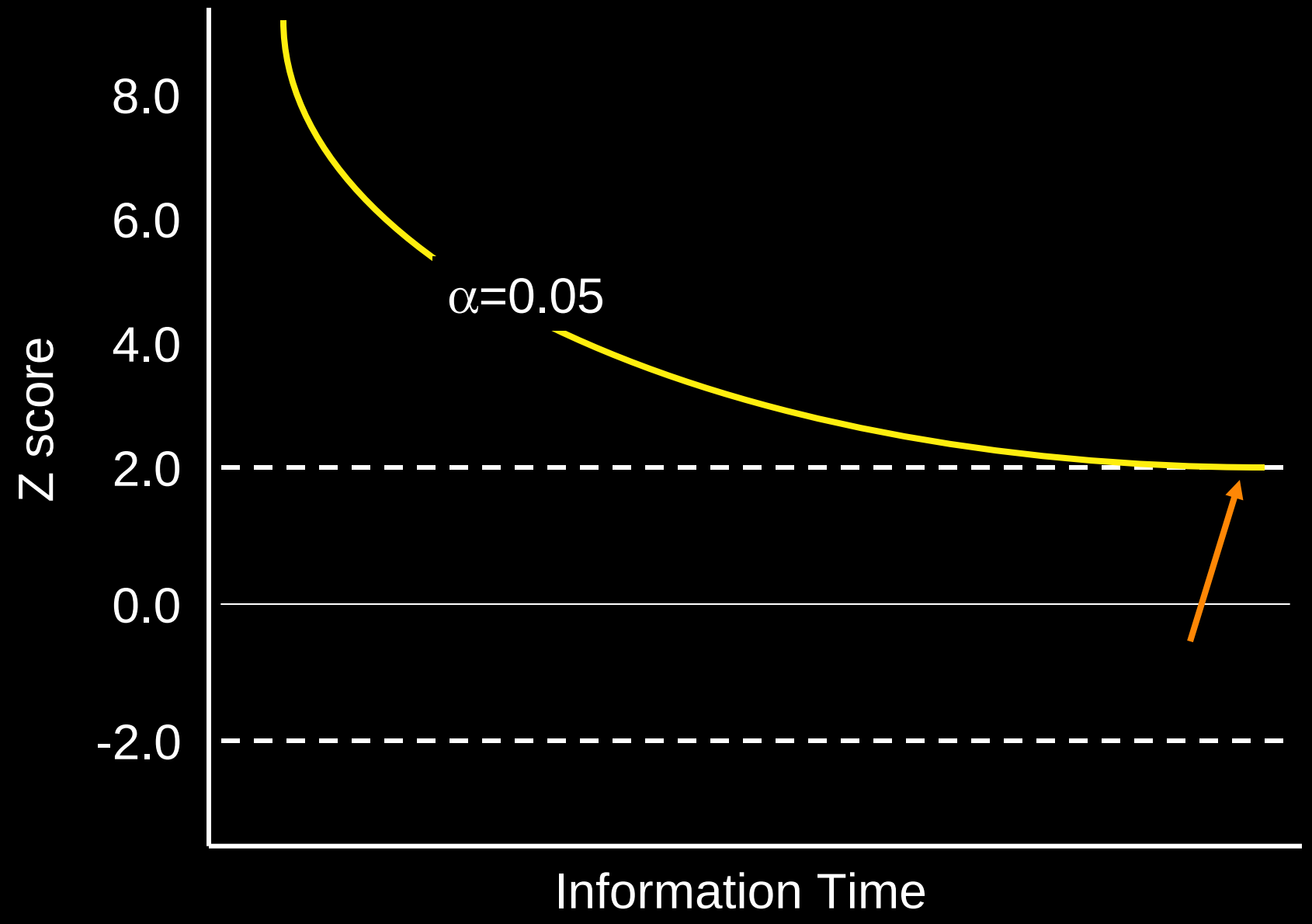
Major Adverse Cardiovascular Event

Placebo	Drug
13	33

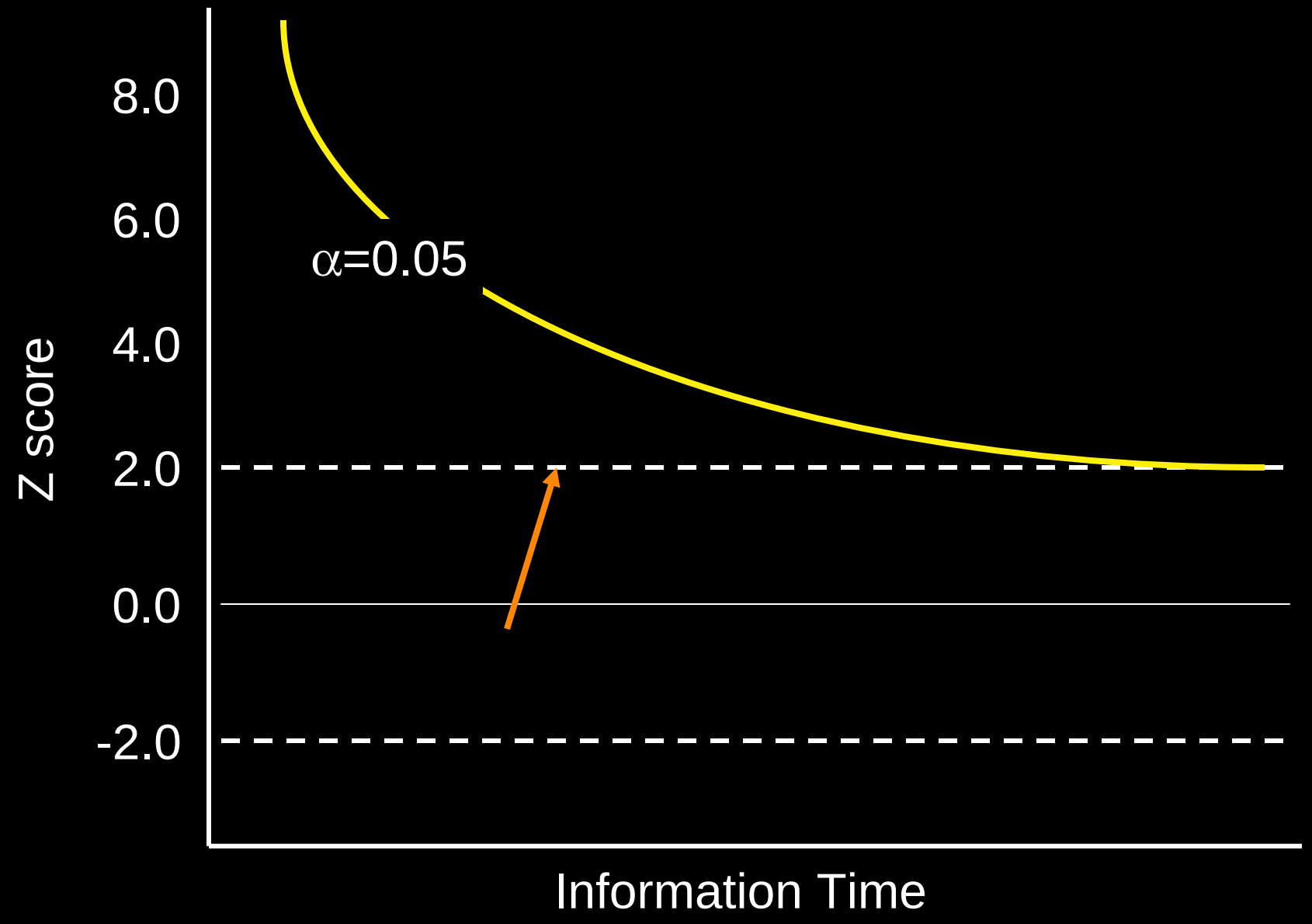
Assumes critical value for α=0.05 is Z=1.96

RR=2.63
95% CI (1.39, 5.00)
P = 0.002

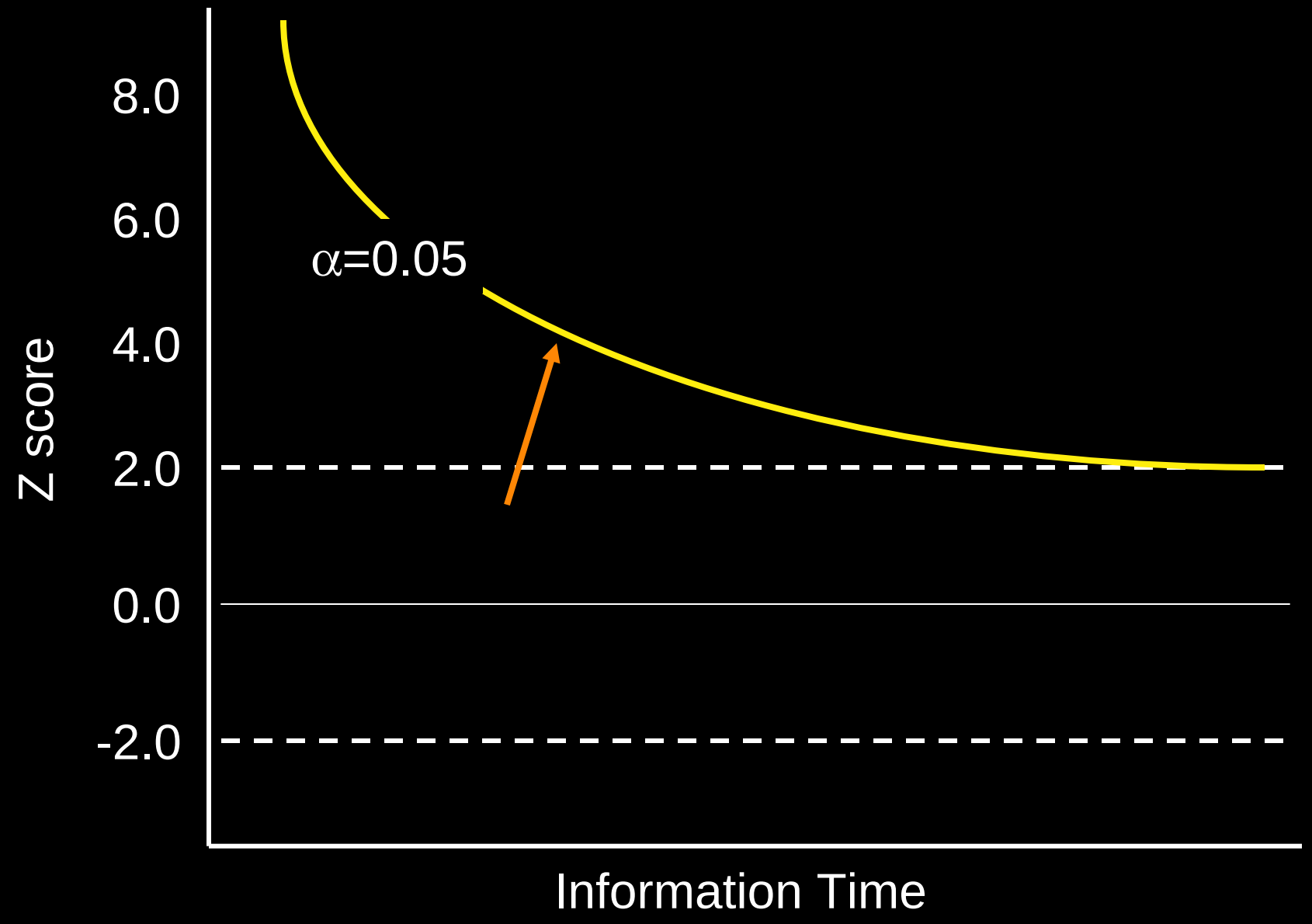
Interim Monitoring in Group Sequential Trials



Interim Monitoring in Group Sequential Trials



Interim Monitoring in Group Sequential Trials



How Should Such a Difference Between Treatment Groups Be Interpreted?

Major Adverse Cardiovascular Event

Placebo	Drug
13	33

If assume critical value for α=0.05 is *boundary* Z score (across range of effect sizes)

RR=2.63
95% CI (< 0.8, > 6.0)
P > 0.10

Boundary-Adjusted Confidence Intervals

- Boundary-adjusted confidence intervals appropriately describe the uncertainty inherent in the analysis of small number of events, thereby reducing the false positive error rate.
- Boundary-adjusted confidence intervals do not provide guidance as to the interpretation of imbalances observed in imprecise data.

What Should We Do With Worrisome Trends in Imprecise Data?

What Should We Do With Worrisome Trends in Imprecise Data?

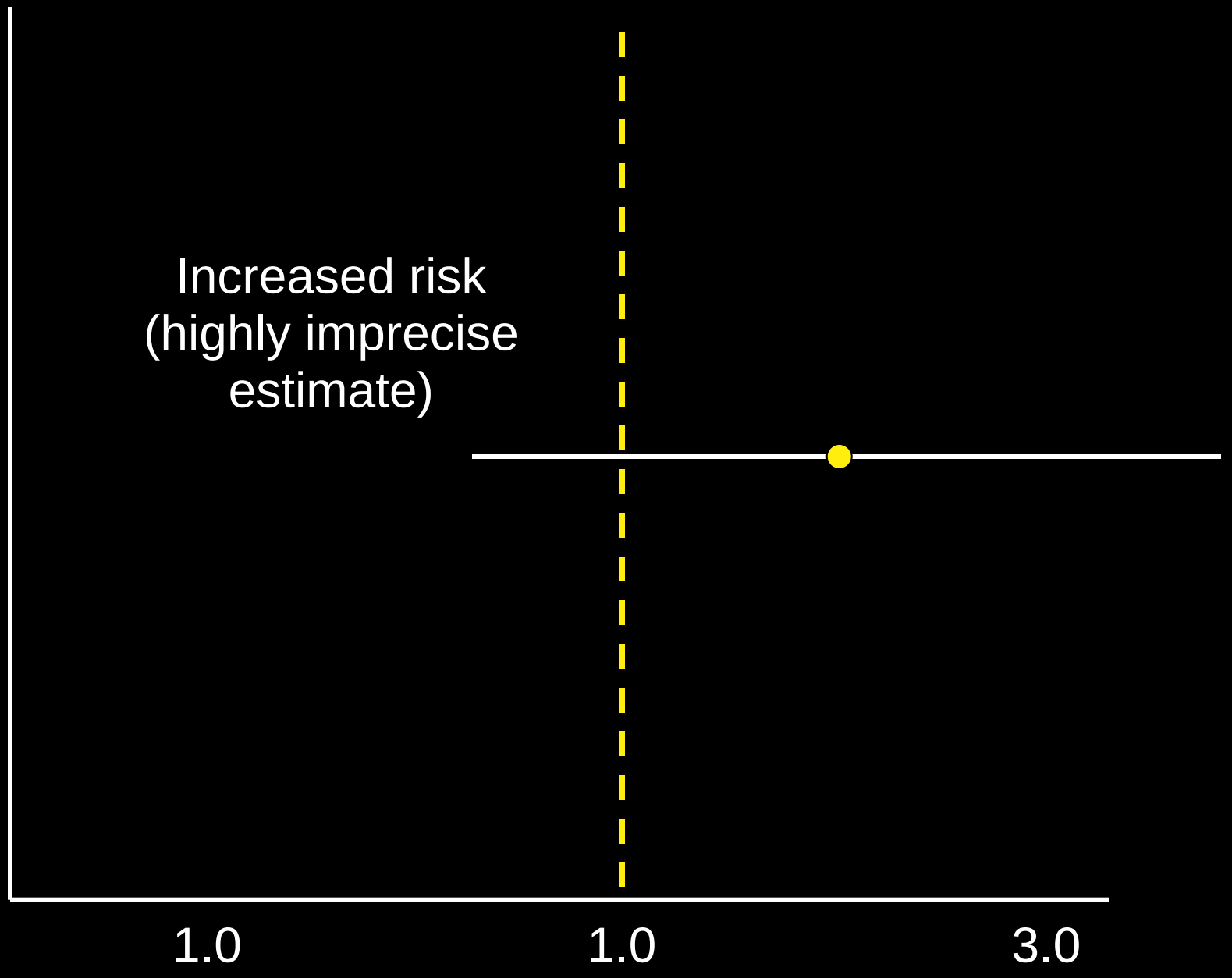
- Believe in observed differences that are biologically plausible.

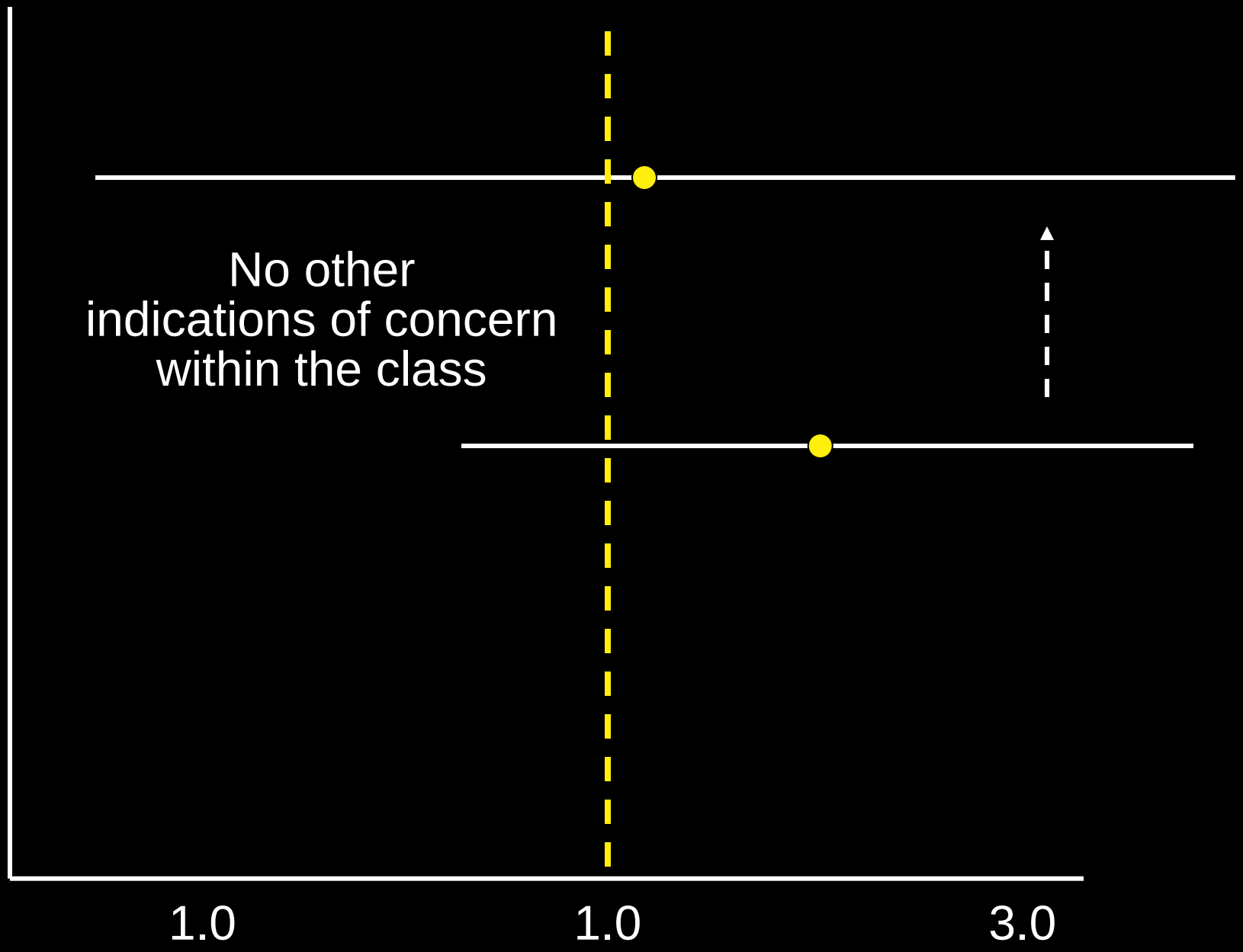
An Indisputable Truth

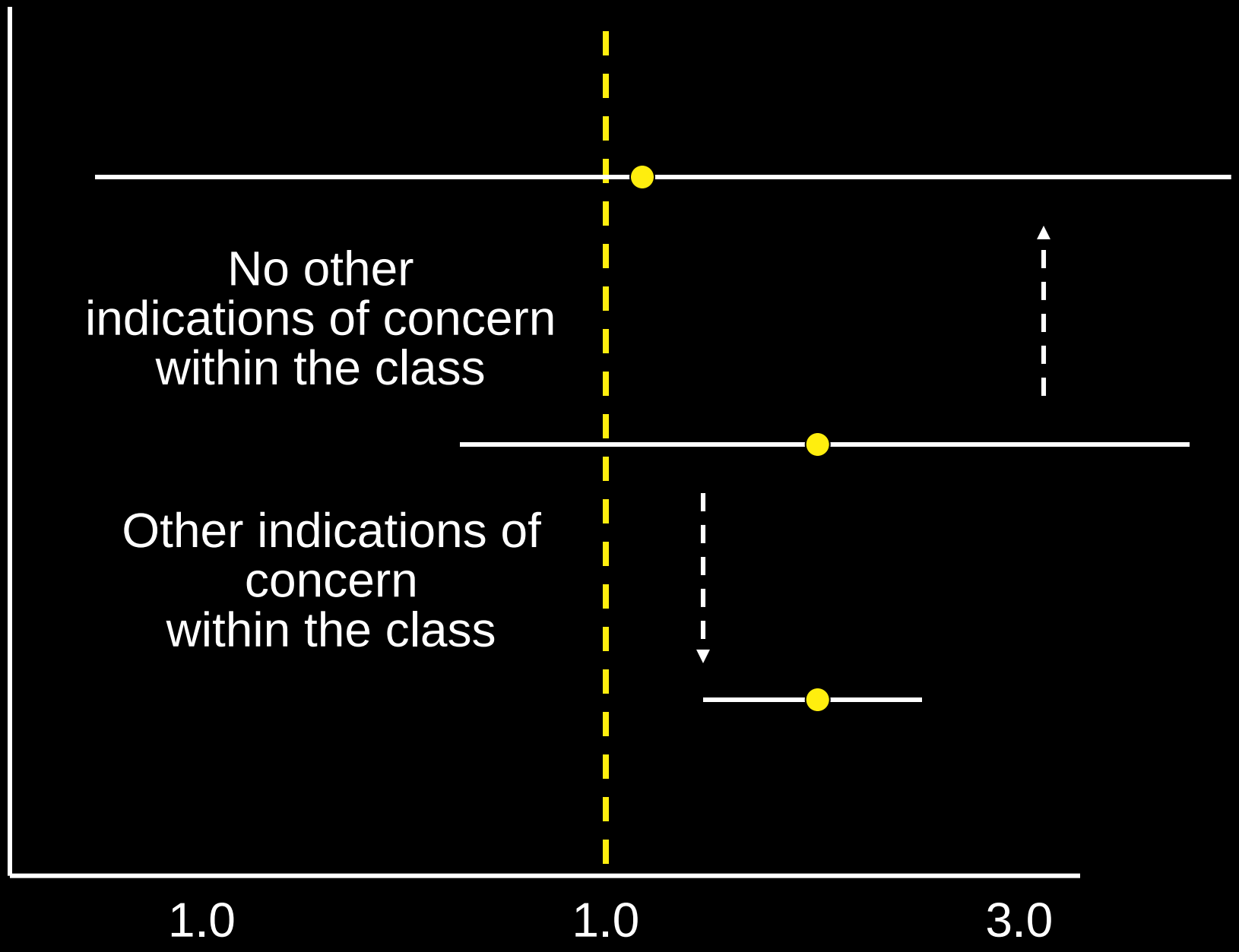
Be wary of differences that are deemed "real" based on biological plausibility, because physicians can always be relied upon to propose a biological mechanism to explain the validity of an unexpected (and potentially preposterous) finding that happens to have an interesting P value.

Reliance on Class Effect

- Interpretation of imbalances in the frequency of adverse events with a new drug can be aided by looking for the presence or absence of similar imbalances in other members of the same drug class.


Increased risk
(highly imprecise
estimate)
1.0
1.0
3.0


No other
indications of concern
within the class
1.0
1.0
3.0


No other indications of concern within the class
Other indications of concern within the class
1.0
1.0
3.0

Reliance on Class Effect

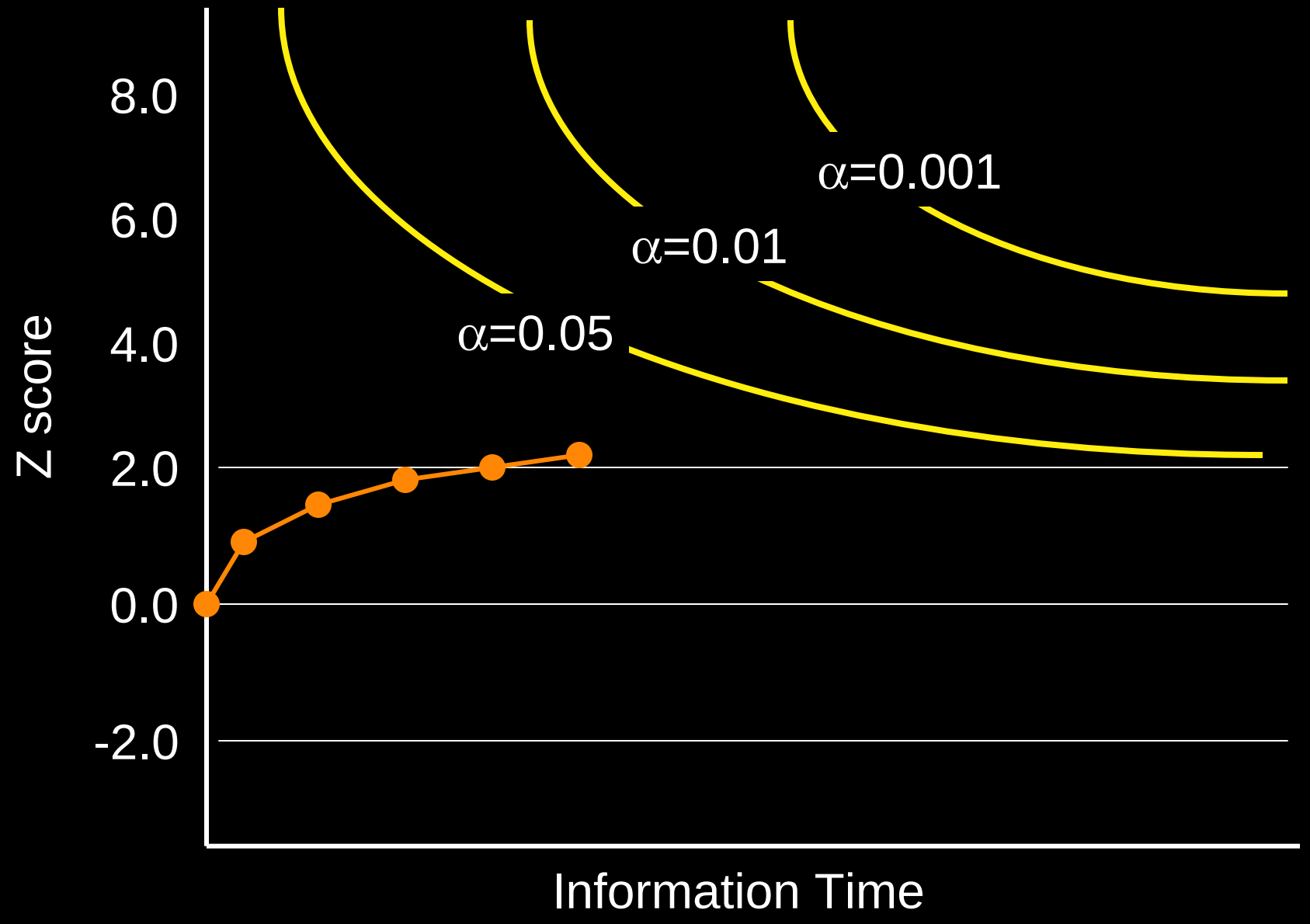
- Interpretation of imbalances in the frequency of adverse events with a new drug can be aided by looking for the presence or absence of similar imbalances in other members of the same drug class.

- Defining and applying the concept of a “drug class” can be difficult.

What Should We Do With Worrisome Trends in Imprecise Data?

- Believe in observed differences that are biologically plausible.
- Look for confirmatory evidence in other studies (avoid being selective).

The Cumulative Meta-Analyses



What Should We Do With Worrisome Trends in Imprecise Data?

- Believe in observed differences that are biologically plausible.
- Look for confirmatory evidence in other studies (avoid being selective).
- Carry out definitive trial with the adverse event as primary endpoint (powered to detect meaningful treatment difference).

Are Definitive Trials the Answer?

- Sponsors pursue *encouraging* trends for important endpoints. [Most are non-confirmatory.] Sponsors should pursue *discouraging* trends for important endpoints. [Most will be non-confirmatory.]
- Definitive trial can address ascertainment and classification biases as well as concerns about multiplicity of comparisons and imprecision of data.

Just Pause and Think

- If you observed an increased frequency of a serious adverse effect in a clinical trial, how easy would you think it would be to carry out a trial intended to definitively evaluate this risk?

Do We Need to Be So Certain When Evaluating Safety Instead of Efficacy?

- We are strict in reaching conclusions about *efficacy* because saying that there is a benefit when there is none means millions will be treated unnecessarily and subject to side effects and costs.
- Some might advocate being less strict in reaching conclusions about *safety*, but saying that there is an adverse effect when there is none means millions will be deprived of an effective treatment.

Conclusions

- The findings of controlled clinical trials are most easily interpreted when they represent the principal efficacy endpoint of the study.
- Safety data is subject to many interpretative difficulties, including ascertainment biases and inflated false positive rates due to the multiplicity of comparisons and imprecision of estimates inherent in analysis of small numbers.
- The FDA, industry and academia remain in a quandary as to how to respond in a responsible fashion to observed differences in reported frequencies of adverse events.

- Return to Main