AptivSolutions℠
Accelerating the Possibilities

The Role of Operating Characteristics in Assesing Bayesian Designs in Pharmaceutical Development

Professor Andy Grieve
SVP Clinical Trials Methodology
Innovation Centre, Aptiv Solutions GmbH
Cologne, Germany

Outline

- Guidelines for Reporting Bayesian Analyses
- Operating Characteristics
- Bayesian Monitoring of Groups Sequential Designs
- Operating Characteristics of Thall and Wathen AD
- Accuracy of Simulation
- Designs Simulation Studies
- Conclusions

Guidelines for Reporting Bayesian Analyses

ROBUST

- Prior Distribution
  - Specified
  - Justified
  - Sensitivity analysis
- Analysis
  - Statistical model
  - Analytical technique
- Results
  - Central tendency
  - SD or Credible Interval

What's Missing?

BAYESWATCH

- Introduction
  - Intervention described
  - Objectives of study
- Methods
  - Design of Study
  - Statistical model
  - Prior / Loss function?
    - When constructed
    - Prior described
    - Loss function described
  - Use of Software – MCMC , starting values, run-in, length of runs, convergence, diagnostics
- Results
- Interpretation
  - Posterior distribution summarized
  - Sensitivity analysis if alternativepriors used

BASIS

- Research question
- Statistical model
  - Likelihood, structure, prior & rationale
- Computation
  - Software - convergence if MCMC, validation, methods for generating posterior summaries
- Model checks, sensitivity analysis
- Posterior Distribution
  - Summaries used: i). Mean, std, quintiles ii) shape of posterior, (iii) joint posterior for multiple comparisons, (iv) Bayes factors
- Results of model checks and sensitivity analyses
- Interpretation of Results
- Limitation of Analysis

What's Missing?
Operating Characteristics

- Type I Error, "Power" etc
- Guidelines written by Bayesians
- Frequentist properties of Bayesian Procedures
  - "Bayesianly Justifiable And Relevant Frequency Calculations For The Applied Statistician" – Don Rubin (1979)
- Objective Bayes – Berger & Bernardo (Uniformative)
- Calibrated Bayes – Rubin, Lewis & Berry, Spiegelhalter
  - Important for pharmaceutical statisticians?

Guidance for the Use of Bayesian Statistics in Medical Device Clinical Trials – FDA/CDRH 2010

- "Because of the inherent flexibility in the design of a Bayesian clinical trial, a thorough evaluation of the operating characteristics should be part of the trial design. This includes evaluation of:
  - probability of erroneously approving an ineffective or unsafe device (type I error)
  - probability of erroneously disapproving a safe and effective device (type II error)
  - power (the converse of type II error: the probability of appropriately approving a safe and effective device)
  - sample size distribution (and expected sample size)
  - prior probability of claims for the device
  - if applicable, probability of stopping at each interim look. "

Operating characteristics: Anti-Bayesian?

- What are “Operating characteristics”?
  - “How often do I make the wrong decision?”
    - Picking the wrong dose to take into later stage testing (Phase III).
    - Deciding to continue developing the drug when the achievable effect size isn’t sufficient.
  - “What if there’s no effect at all?”
    - Type I error
  - “Given that there’s something to find, how soon do I find it, and how often do I get it right?”
    - ASN
    - Time to run study
    - Type II error
- Anti-Bayesian?

Operating characteristics

- Operating characteristics aim to answer many types of queestions
  - If there is no effect, how often do we find a spurious effect?
  - If there is an effect, how often do we pick it up?
  - Do we make the right choice of dose?
  - If we are estimating something, how close do we get to the “true” value?

Simulation Results – Bayesian Adaptive Design (NI)
Nitin Patel(Cytel) – FDA/DIA Stats Forum (Apr 2011)
10^5 simualtions

Treatment Effect (P_T)	Power	Fixed Sample Size	SSA+ES Sample Size	Trial Duration (weeks)
0.145	0.025	NA	408	103
0.245	0.739	418	372	87
0.275	0.911	418	345	77
0.305	0.979	419	319	68
0.335	0.997	426	299	62

For fixed design with sample size of 418 the trial duration is 108 weeks

Expected improvement when $P_T = 0.275$: 17% fewer patients, 31 wks shorter

Bayesian Monitoring of Clinical Trials

- Development follows Grossman et al (SIM, 1994)
  - All data & priors are normal (known variance σ^2)
  - A maximum of n patients in each of 2 groups (trts:A and B)
  - T interim analyses after tn/T (t=1,..T) patients per group
  - Of interest is $\delta=\mu_A-\mu_B$
  - The observed difference between groups of the t^{th} cohort is d_t with variance $T\sigma_\delta^2/n$ (where $\sigma_\delta^2=2\sigma^2$)
  - Prior information for δ is available: corresponding to fn patients per group centred at δ_0
- Bayes Theorem implies that at the t^{th} interim the posterior for δ is:

$$p\left(\delta \mid D_t = \sum_{i=1}^{t} d_i\right) \sim N\left(\frac{\frac{n}{T}D_t + fn\delta_0}{\frac{tn}{T} + fn}, \frac{\sigma^2}{\frac{tn}{T} + fn}\right)$$

Bayesian Monitoring of Clinical Trials

- Stopping rule: $\mathrm{Prob}(\delta > \delta_C \mid D_t) > 1 - \psi_t$

  equivalent to: $\Psi_t > \Phi\left[\dfrac{\delta_C - (D_t + Tf\delta_0)/(t + fT)}{\sigma/(tn/T + fn)^{1/2}}\right]$

  requiring

$$D_t > -\frac{T^{1/2}(t + fT)^{1/2} Z_{\Psi t}\sigma}{n^{1/2}} + \delta_C(t + fT) - Tf\delta_0$$

- This is the general case and there are a number of "tuning" parameters: ψ_t, f, δ_C and δ_0

Bayesian Monitoring of Clinical Trials
Special Case 1: T=1, δ_C=0

- Stopping rule requires: $D > -\frac{(1+f)^{1/2} Z_\psi \sigma}{n^{1/2}} - f\delta_0$

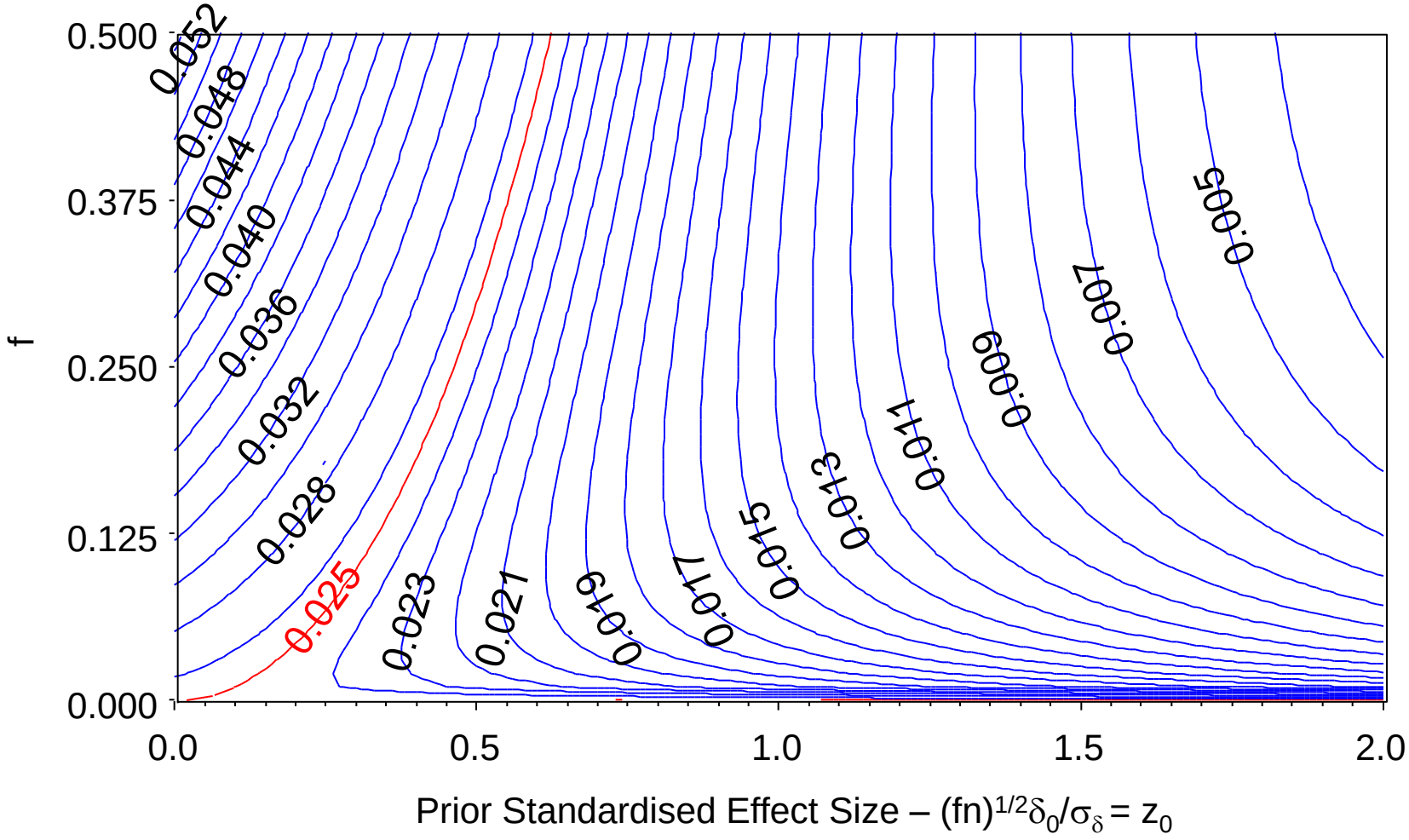
- What are the frequency properties of this rule?
- Under the Null Hypothesis: $\delta \sim N(0,\sigma^2_\delta/n)$

$$\Rightarrow P\left[D > \frac{-\sigma_\delta z_\psi (1+f)^{1/2}}{n^{1/2}} - f\delta_0\right] = 1 - \Phi\left(-z_\psi (1+f)^{1/2} - \frac{f^{1/2}(nf)^{1/2}\delta_0}{\sigma_\delta}\right)$$

- To control this at the 2.5% level we need

$$z_{1-\psi} = \frac{z_{0.975} + f^{1/2} z_0}{(1+f)^{1/2}}$$

Contours of Bayesian Decision Rule (ψ) To give a One-sided Type I Error 0f 2.5%



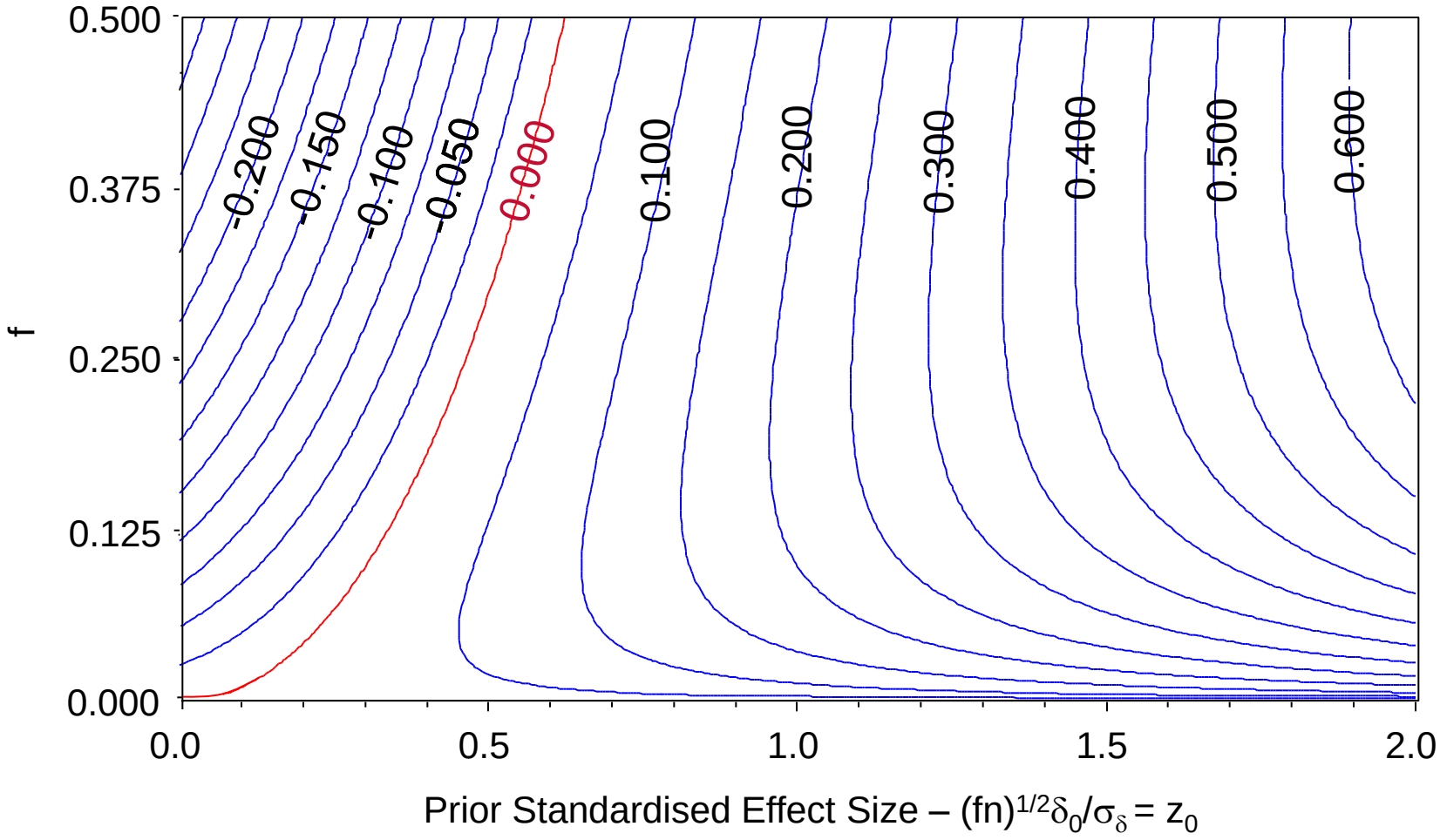
Bayesian Monitoring of Clinical Trials
Special Case 2: T=1, ψ=0.025

- In this case

$$\mathrm{Prob}(\delta > \delta_c \mid D) = 0.975 = 1 - \Phi\left(\frac{n^{1/2}\left[(1+f)\delta_c - (D + f\delta_0)\right]}{\sigma_\delta (1+f)^{1/2}}\right)$$

giving a condition for D which can be used to find a value for δ_c to give the appropriate type I error.

Contours of Bayesian Decision Rule ($\delta_C\sigma_\delta/n^{1/2}$) To give a One-sided Type I Error 0f 2.5%



Bayesian Monitoring of Clinical Trials
Special Case 1 & 2

- Whichever approach is used is turns out that using this approach is effectively discounting the prior information.
- To see this substitute $z_\psi = \dfrac{z_{0.025} - f^{1/2} z_0}{(1+f)^{1/2}}$ into

$$D > -\frac{(1+f)^{1/2} Z_\psi \sigma}{n^{1/2}} - f\delta_0$$

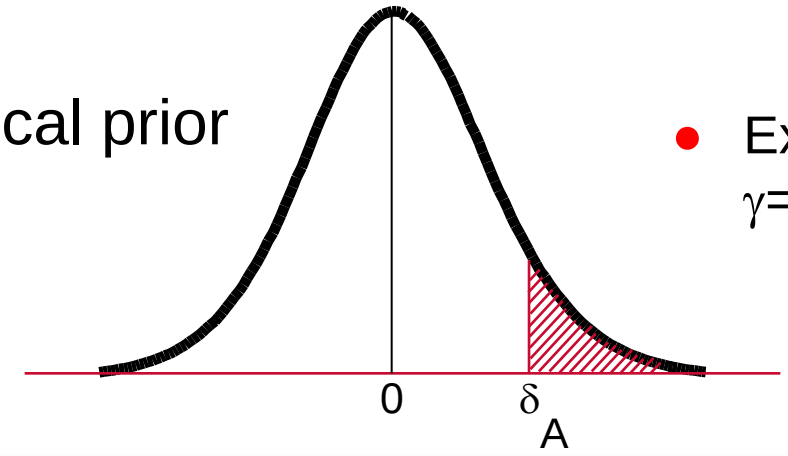
to give

$$D > \frac{\sigma_\delta z_{0.975}}{n^{1/2}}$$

the standard, frequentist decision criteria – 100% discounting

Bayesian Monitoring of Clinical Trials
Special Case 3: ψ_t=0.025, δ_C=0, δ_0=0 (Sceptical Prior)

- A sceptical prior can be set up formally.
- Prior centred around 0, with a small probability g of achieving the alternative (δ_A) - $p(\delta>\delta_A) = \gamma$
- From which $\delta_A = -\dfrac{\sigma_\delta z_{1-\gamma}}{(fn)^{1/2}}$



- Now suppose the trial has been designed with size α and power 1-β to detect the alternative hypothesis δ_A.
- So that

$$n = \frac{\sigma_\delta^2 \left(z_{1-\alpha/2} + z_{1-\beta}\right)^2}{\delta_A^2}$$

- From which: $f = \left(\dfrac{z_\gamma}{z_{1-\alpha/2} + z_{1-\beta}} \right)^2$
- Example: α=0.05, 1- β=0.90, γ=0.05 => f ~ 1/4

Bayesian Monitoring of Clinical Trials
Special Case 3: ψ_t=0.025, δ_C=0, δ_0=0

- In this case: $$\mathrm{Prob}(\delta > \delta_C \mid D_t) = 1 - \Phi\left[\frac{-D_t/(t+fT)}{\sigma_\delta/(t/T+f)^{1/2}}\right] > 0.975$$

$$= \Phi\left[\frac{T^{1/2}D_t}{\sigma_\delta n^{1/2}(t+fT)^{1/2}}\right] > 0.975$$

$$= \Phi\left[\frac{T^{1/2}D_t}{\sigma_\delta (nt)^{1/2}}\frac{t^{1/2}}{(t+fT)^{1/2}}\right] > 0.975$$

which is equivalent to increasing the critical region by a factor

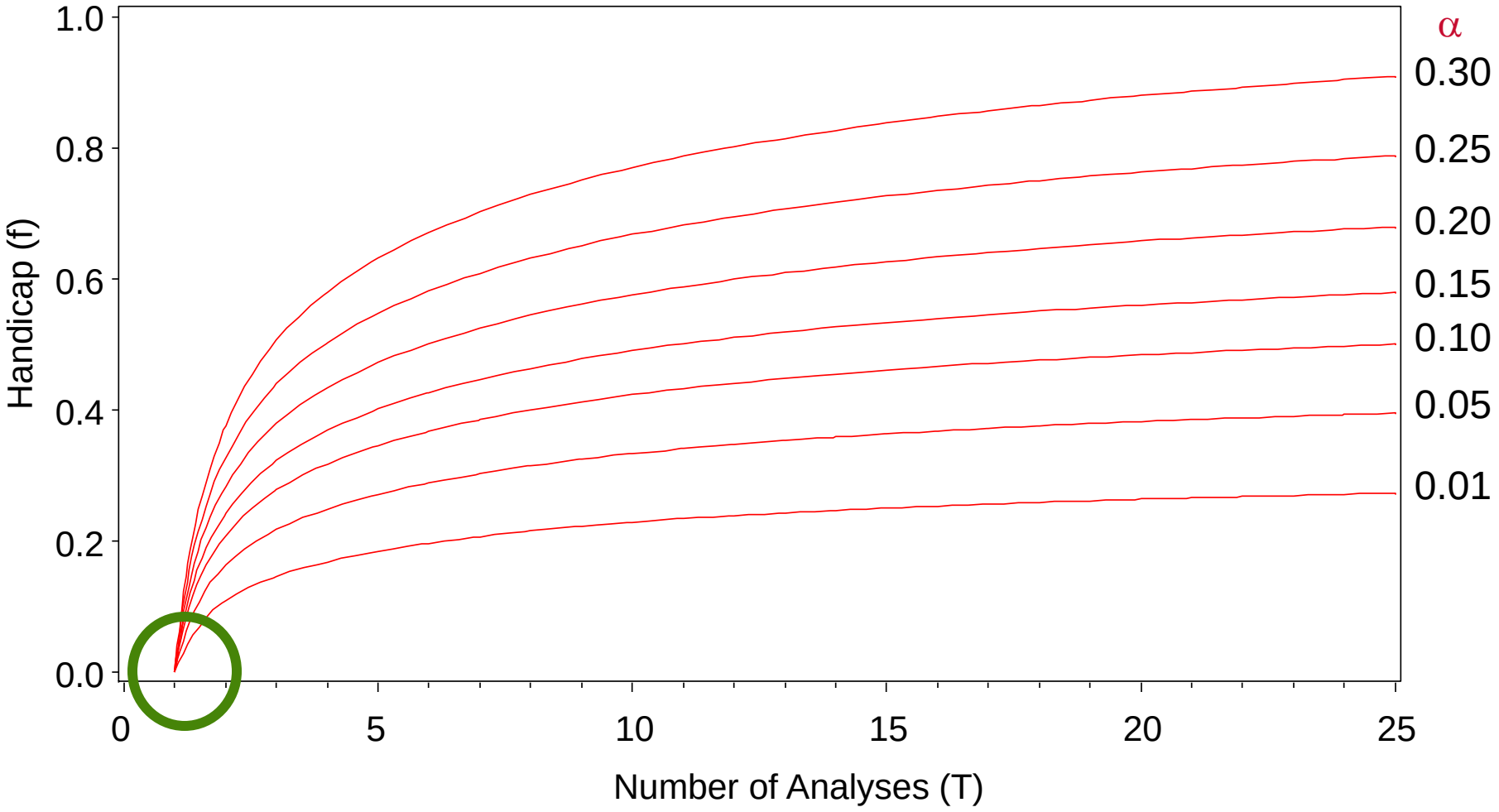
$$\sqrt{\frac{t+fT}{t}}$$

Grossman et al(1994) call f the "handicap"

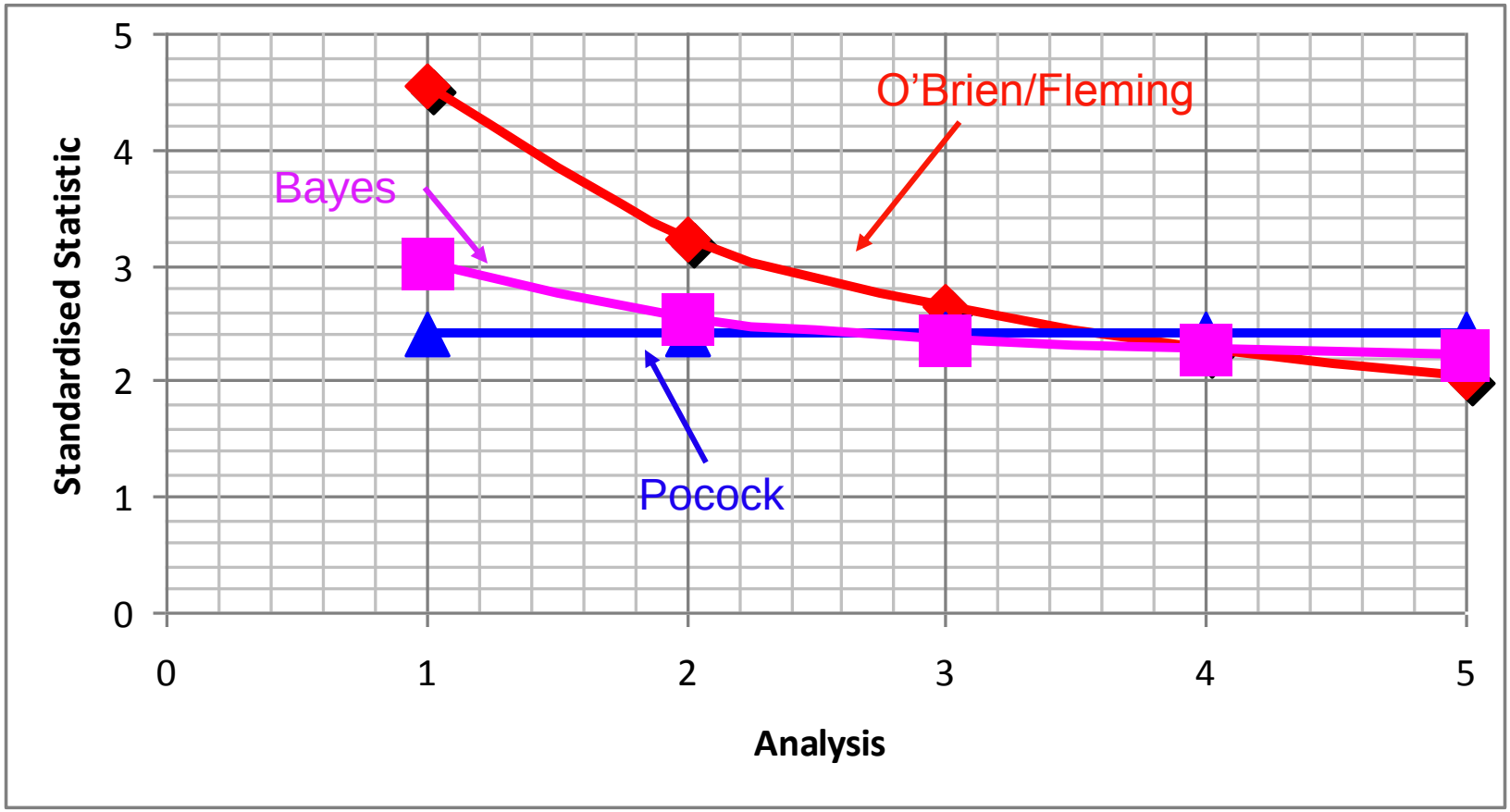
Bayesian Monitoring of Clinical Trials Special Case 3: $\psi_t=0.025$, $\delta_C=0$, $\delta_0=0$

- The frequentist properties of this handicapping are not so easy to derive.
- For T=2 – a single interim – the frequentist type-I error can be calculated using a bivariate normal probability function, e.g. the SAS function PROBNRM.
- For T > 3 Grossman et al (1994) use simulation to determine the handicap f that controls the two-sided type I error at 5% and 1% (20,000,000 trials)
- Alternatively use can be made of the algorithm derived by Armitage, MacPherson and Rowe (JRSSA, 1969) – used a SAS implementation of FORTRAN program by Reboussin, DeMets, Kim and Lan or SEQ, SEQSCALE & SEQSHIFT (PROC IML)

Handicaps(f) To Control the Two-sided α for Upto 25 Analyses



Comparison of Critical Values
O'Brien/Fleming, Pocock & Handicapped Bayes



Handicapped Bayes versus Optimal Designs (Pocock, 1982)

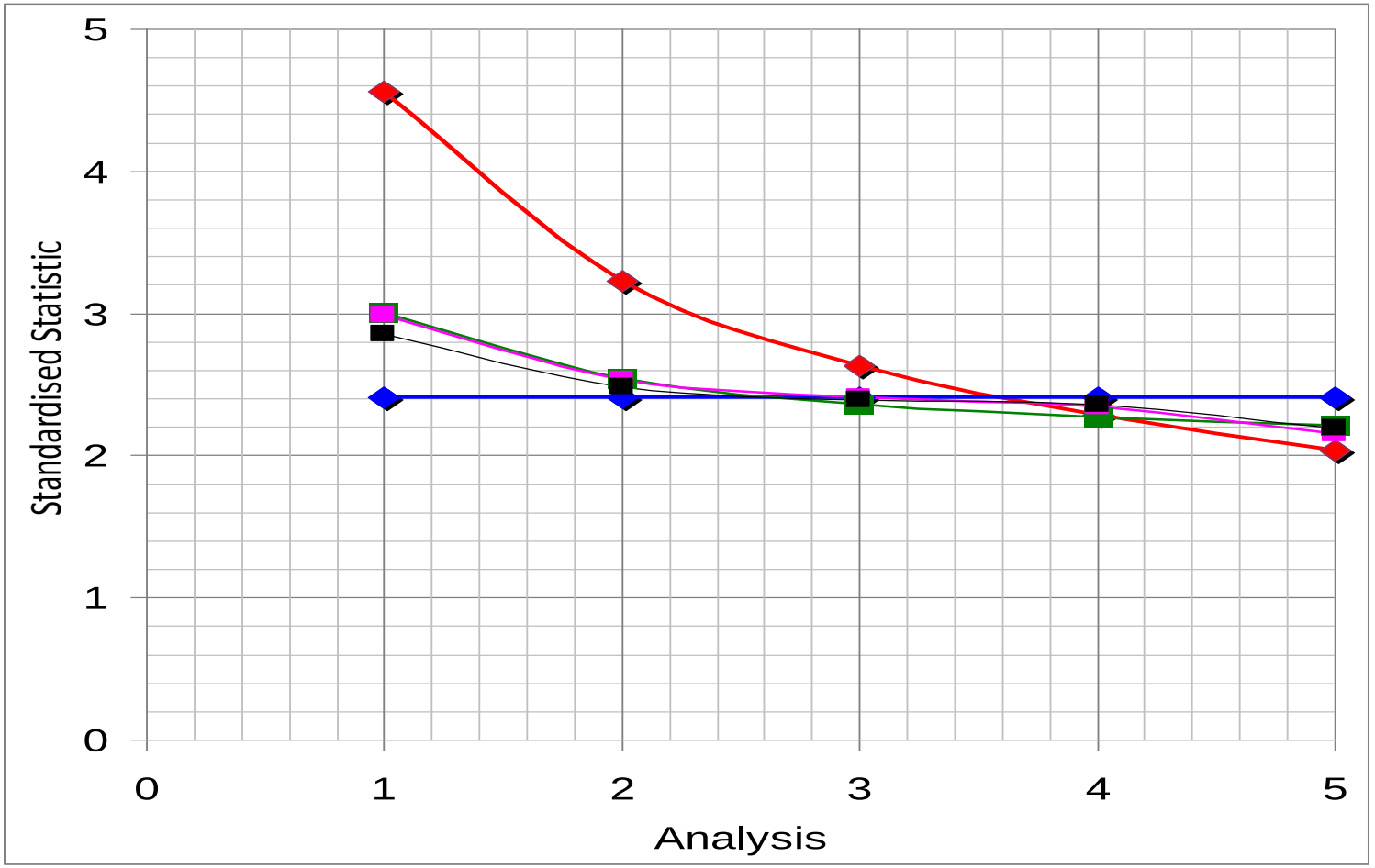
- Investigated properties of group sequential designs, in particular the Average Sample Number (ASN)

Maximum number of groups, K	Nominal significance level, α'	Required number* of patients per group 2n	Maximum number* of patients 2nN	Average number of patients until stopped under H_A (ASN)
1	0.05	51.98	52.0	52.0
2	0.0294	28.39	56.8	37.2
3	0.0221	19.73	59.2	33.7
4	0.0182	15.19	60.8	32.3
5	0.0158	12.38	61.9	31.3
10	0.0106	6.50	65.0	29.8
20	0.0075	3.38	67.6	29.5

Multiply by σ^2/δ^2

- Generated “optimal” GSDs, for given power minimum ASN

Comparison of Critical Values
Optimal (75% & 80% Power) & Handicapped Bayes

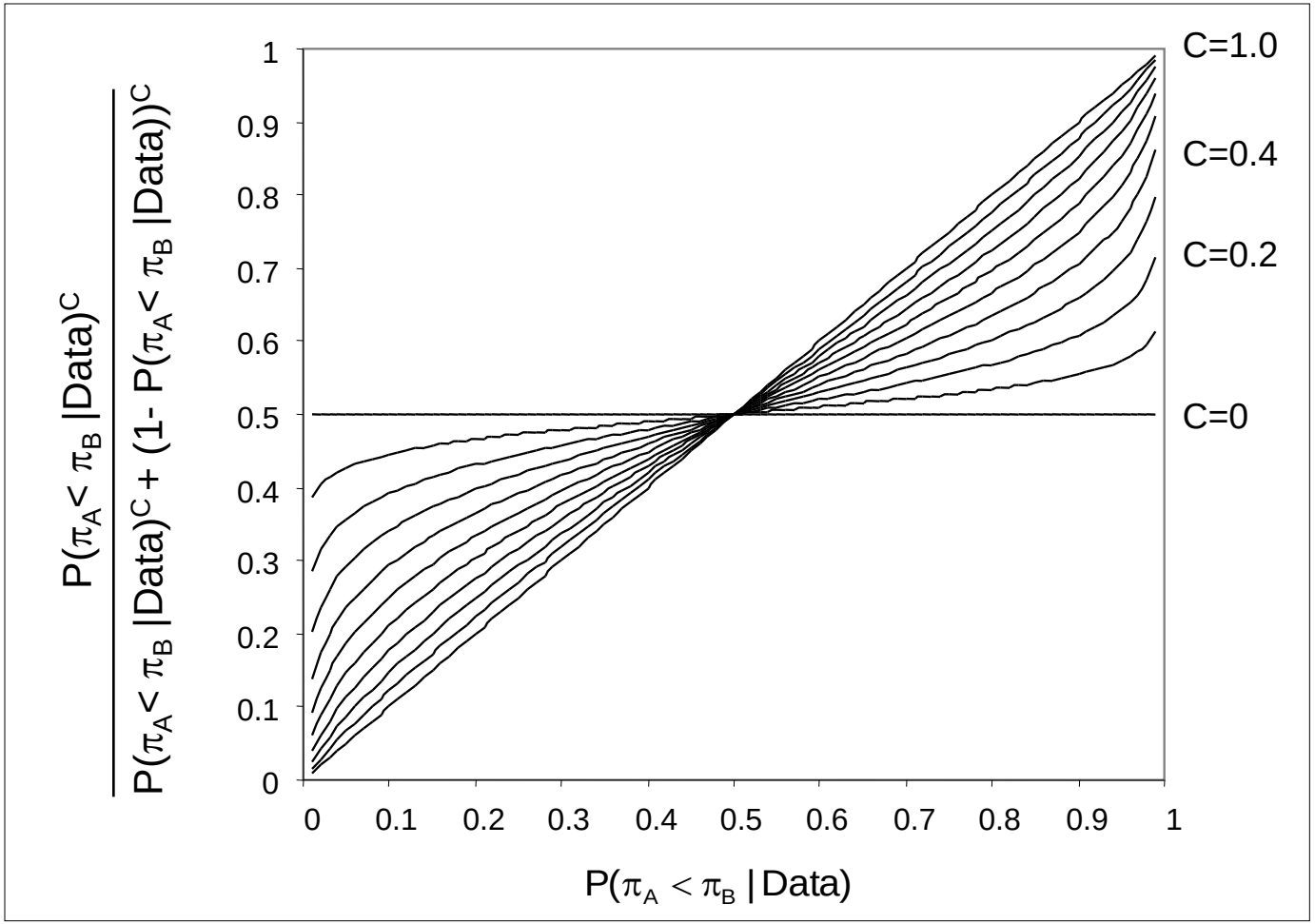


Bayesian Adaptive Randomisation
Thall and Wathen (Eur J Cancer, 2007)

- Back to the idea of Thompson (1933)
- Similar to RPW – binary outcome
- Randomisation to treatment A on the basis of a function of $P(\pi_A < \pi_B|\text{Data})$ although in practice Thompson used $P(\pi_A < \pi_B|\text{Data})$.
- Unstable
- Thall and Wathen (2007)

$$\frac{P(\pi_A < \pi_B \mid \text{Data})^C}{P(\pi_A < \pi_B \mid \text{Data})^C + \left(1 - P(\pi_A < \pi_B \mid \text{Data})\right)^C}$$

Bayesian Adaptive Randomisation
Impact of Choice of C



Bayesian Adaptive Randomisation
Impact of Choice of C

- Thall and Whalen recommend C= n/(2N)
  - n=current sample size
  - N=study's maximum sample size
- Begins with C=0, ends with C=1/2
- C=1/2 "works well in many applications"

- Giles et al (J Clin Oncology, 2003)
  - Similar idea – but now with 3 arms (2 experimental, 1 control) using functions of $P(m_1<m_0|data)$, $P(m_2<m_0|data)$, and $P(m_1<m_2|data)$, - m_2, m_1, and m_0 are the median survival times

2 x 2 Contingency Table

- Data structure

	Response	No Response
Treatment A	n_{11} (π_A)	n_{12} (1-π_B)
Treatment B	n_{21} (π_B)	n_{22} (1-π_B)

Likelihood $$\pi_A^{n_{11}}\left(1-\pi_{1A}\right)^{n_{12}}\pi_B^{n_{21}}\left(1-\pi_B\right)^{n_{22}}$$

Prior $$\propto \pi_A^{\nu_{11}-1}\left(1-\pi_A\right)^{\nu_{12}-1}\pi_B^{\nu_{21}-1}\left(1-\pi_B\right)^{\nu_{22}-1}$$

Posterior $$\propto \pi_A^{n_{11}+\nu_{11}-1}\left(1-\pi_A\right)^{n_{12}+\nu_{12}-1}\pi_B^{n_{21}+\nu_{21}-1}\left(1-\pi_B\right)^{n_{22}+\nu_{22}-1}$$

2x2 Contingency Table - Posterior Inference

- The probability of interest is

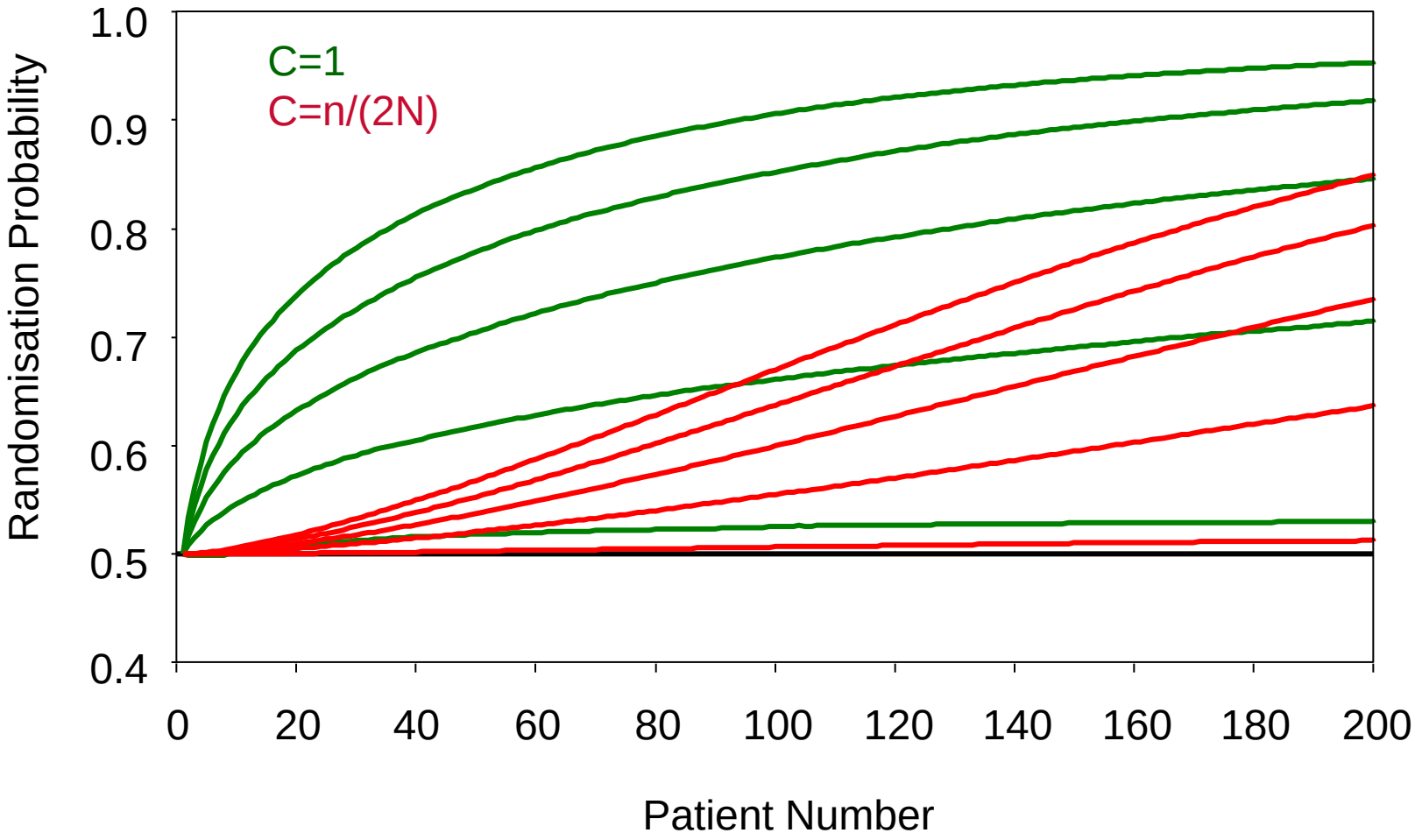
$$\mathrm{Prob}(\pi_A < \pi_B \mid \mathrm{Data}) = \sum_{k=\max(n_{21}+\nu_{21}-n_{12}-\nu_{12},0)}^{n_{21}+\nu_{21}-1} \frac{\binom{n_{11}+\nu_{11}+n_{21}+\nu_{21}-1}{k}\binom{n_{12}+\nu_{12}+n_{22}+\nu_{22}-1}{n_{21}+\nu_{21}+n_{22}+\nu_{22}-1-k}}{\binom{n_{11}+\nu_{11}+n_{21}+\nu_{21}+n_{12}+\nu_{12}+n_{22}+\nu_{22}-2}{n_{11}+\nu_{11}+n_{12}+\nu_{12}-1}}$$

based on the cumulative hypergeometric function as is Fisher's exact test:

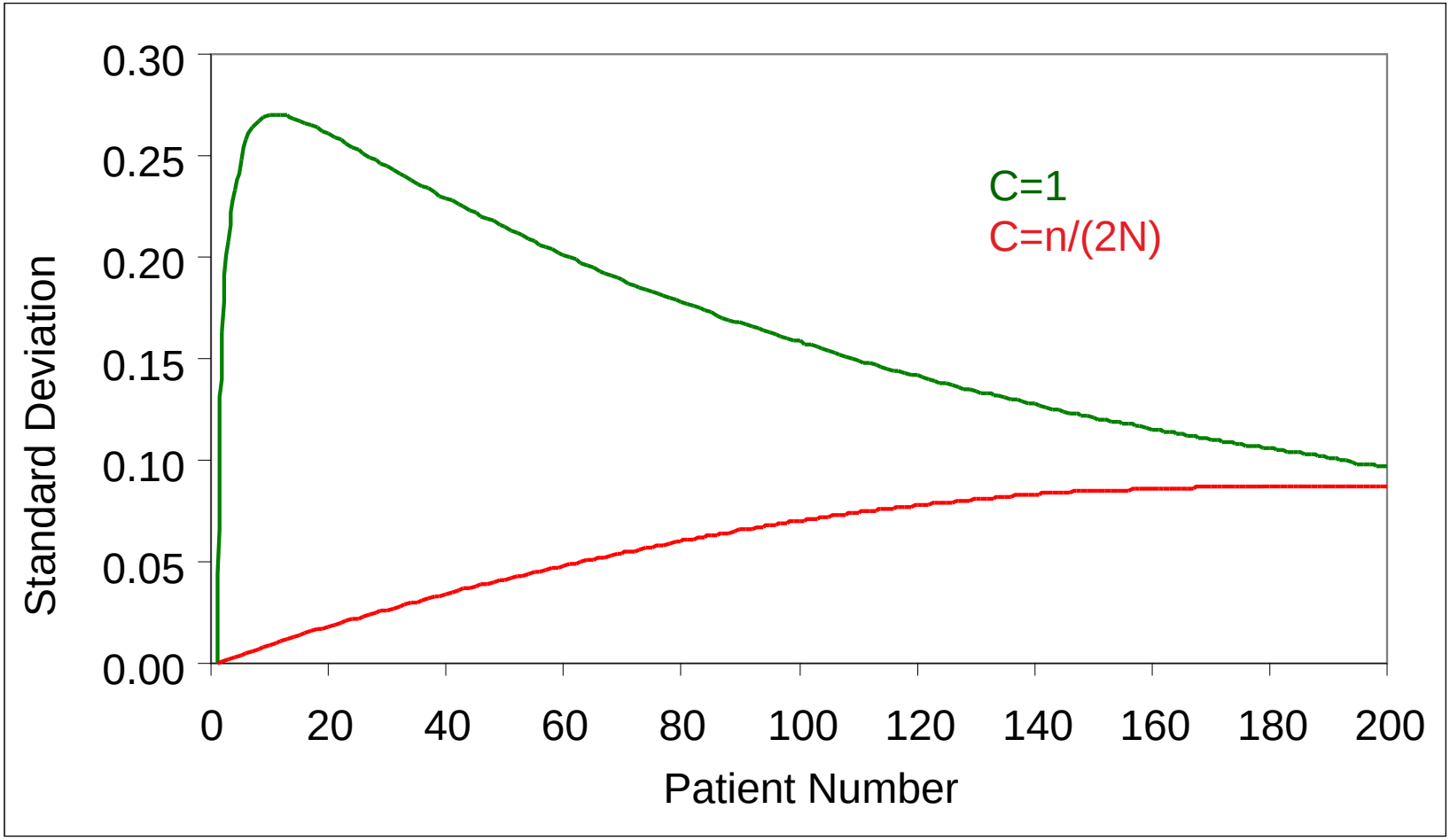
Bayesian AD – Thall & Wathen(EJC,2007) Type-I Error Based on T&W Criterion

- Thall & Wathen illustration is based on:
  - N = 200
  - Stopping Rules
    - If $P(\pi_A<\pi_B|Data) > 0.99$ stop and "choose" B
    - If $P(\pi_A<\pi_B|Data) < 0.01$ stop and "choose" A (futility)
- What does the type I error look like ?
- A complication is that the control rate, π_A, is a nuisance parameter

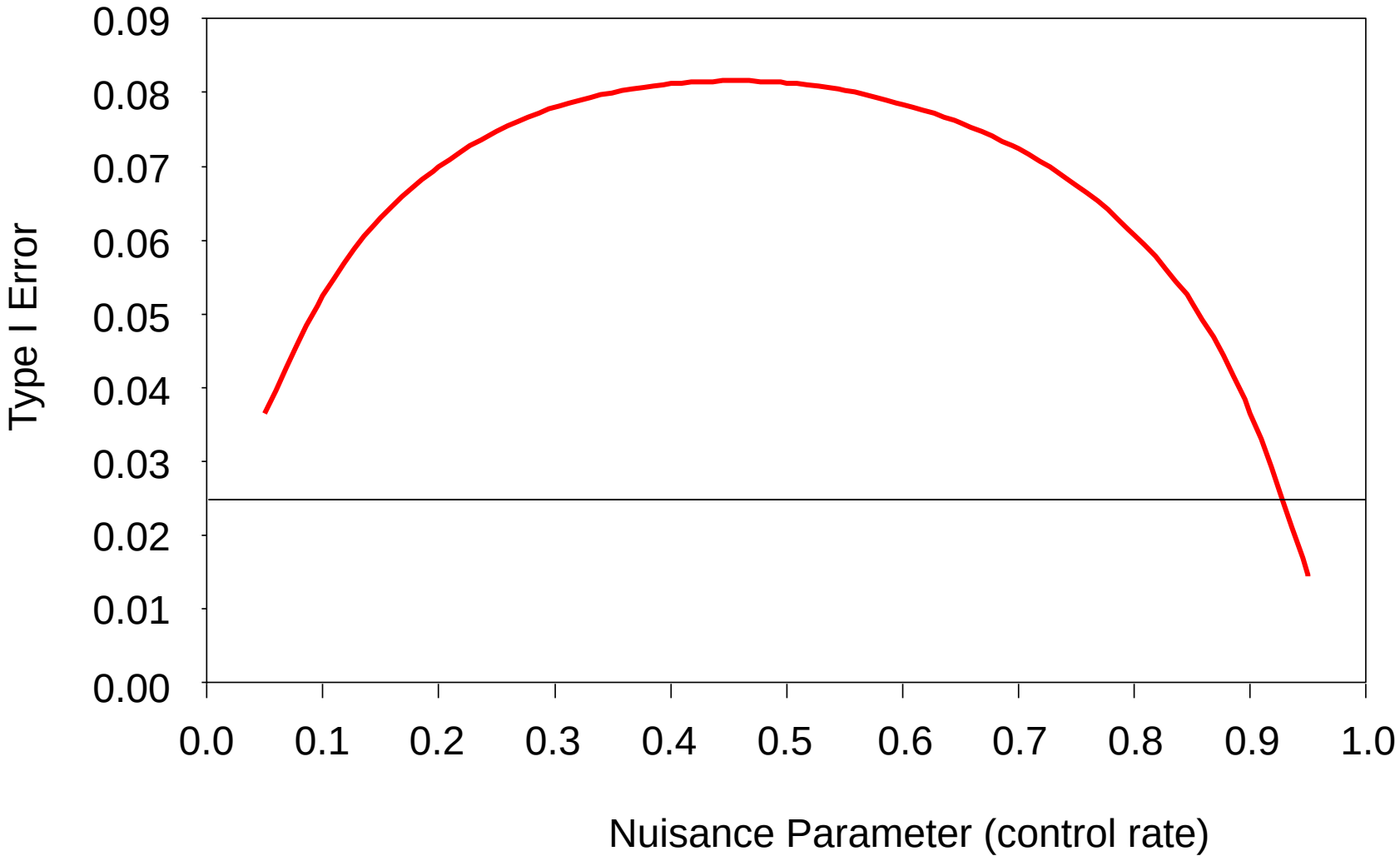
Bayesian AD – Thall & Wathen(EJC,2007) N=200
Randomisation Probabilities (10^5 simulations)
π_A=0.25 , π_B=0.25(0.05)0.45



Bayesian AD – Thall & Wathen(EJC,2007) N=200
Variability of Randomisation Probabilities
π_A=0.25 , π_B=4.25



Bayesian AD – Thall & Wathen(EJC,2007) N=200
Type-I Error Based on T&W Criterion - $P(\pi_A > \pi_B | Data) > 0.99$
10^6 Simulations / control rate



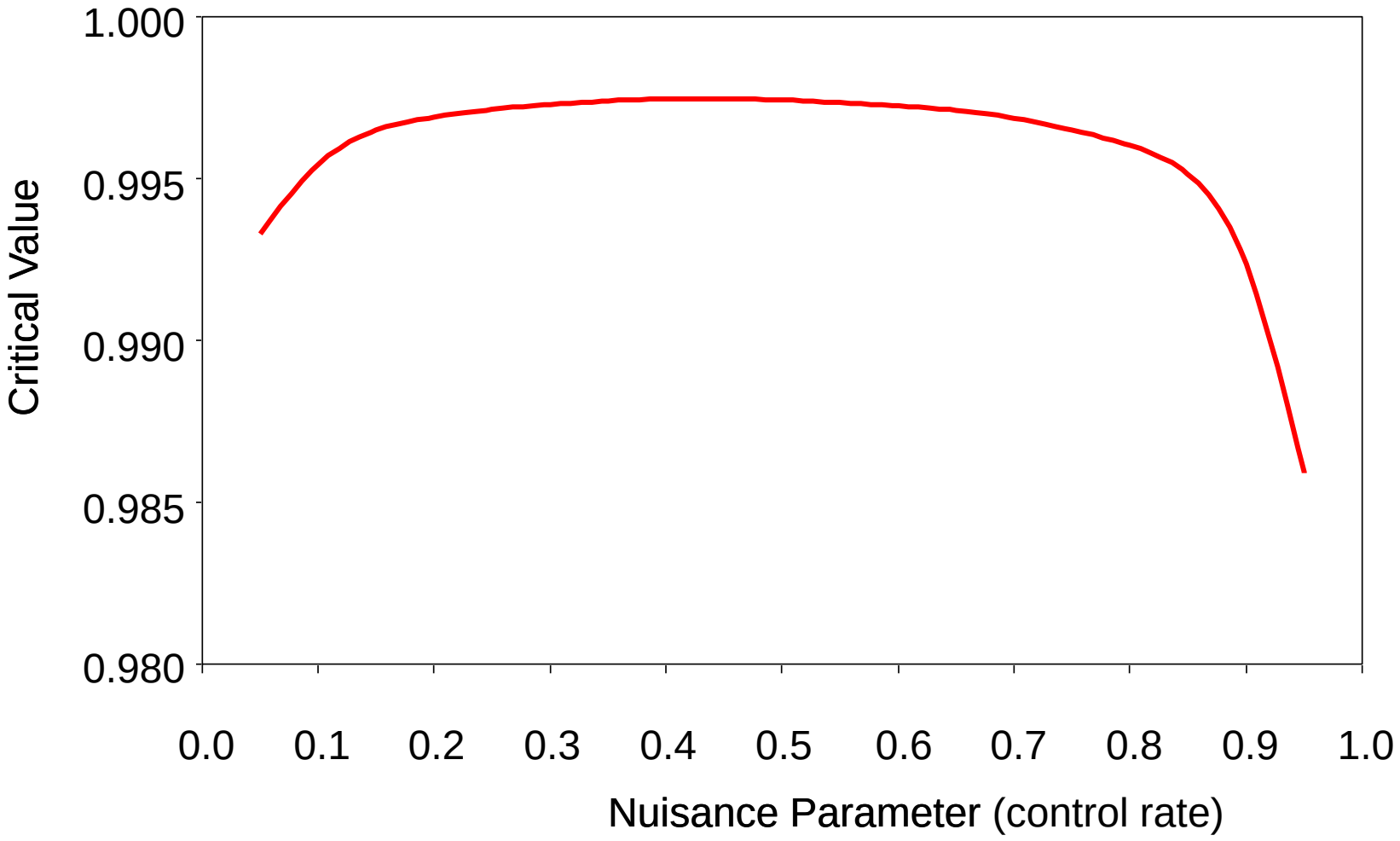
Bayesian AD – Thall & Wathen(EJC, 2007) N=200 Control One-Sided Type-I Error

- The issue is the number of tests being conducted

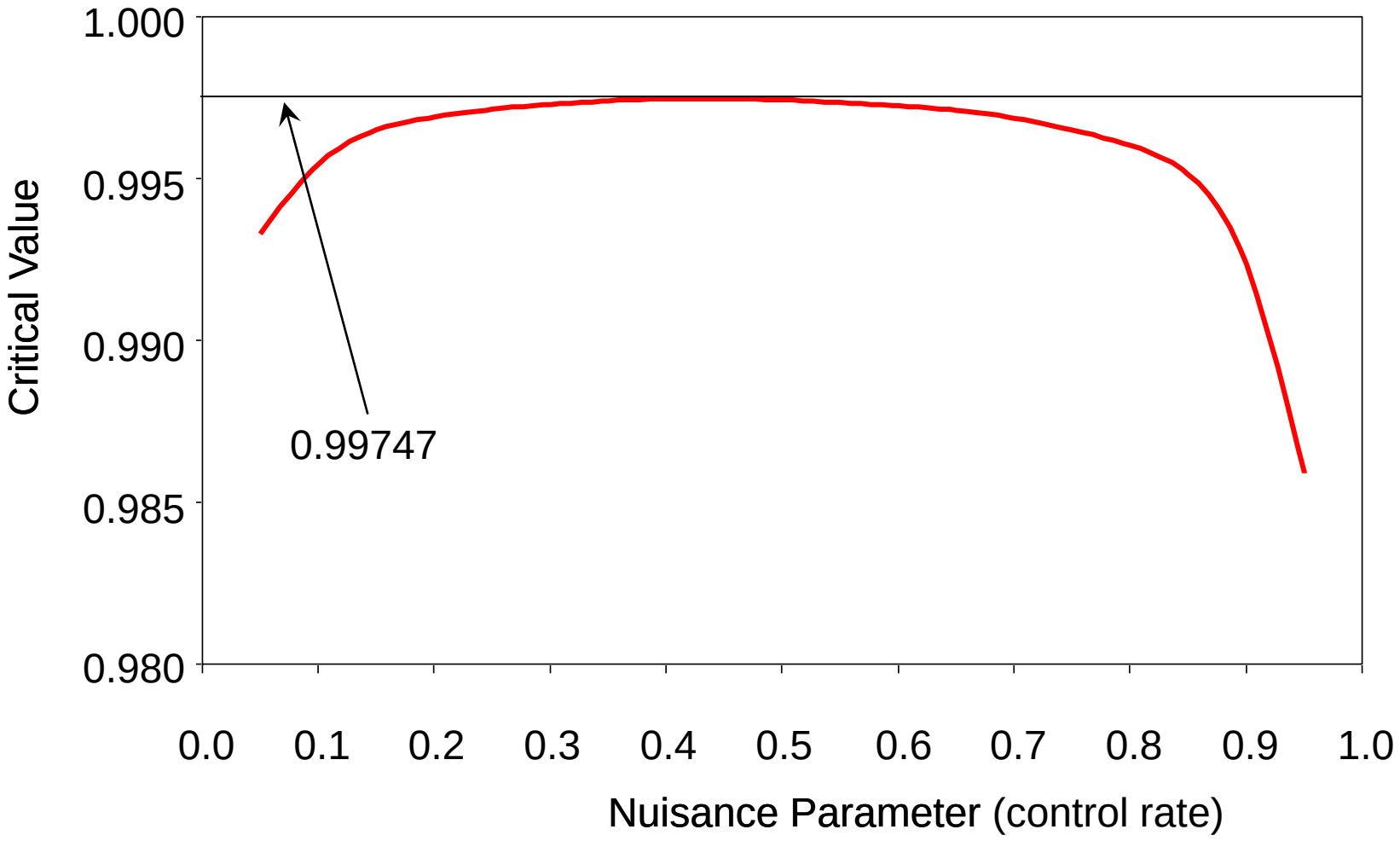
1. Reduce the problem using cohorts (20, 50 ?)

2. Or choose decision criterion $P(\pi_A<\pi_B|Data)>$? to control type-I error

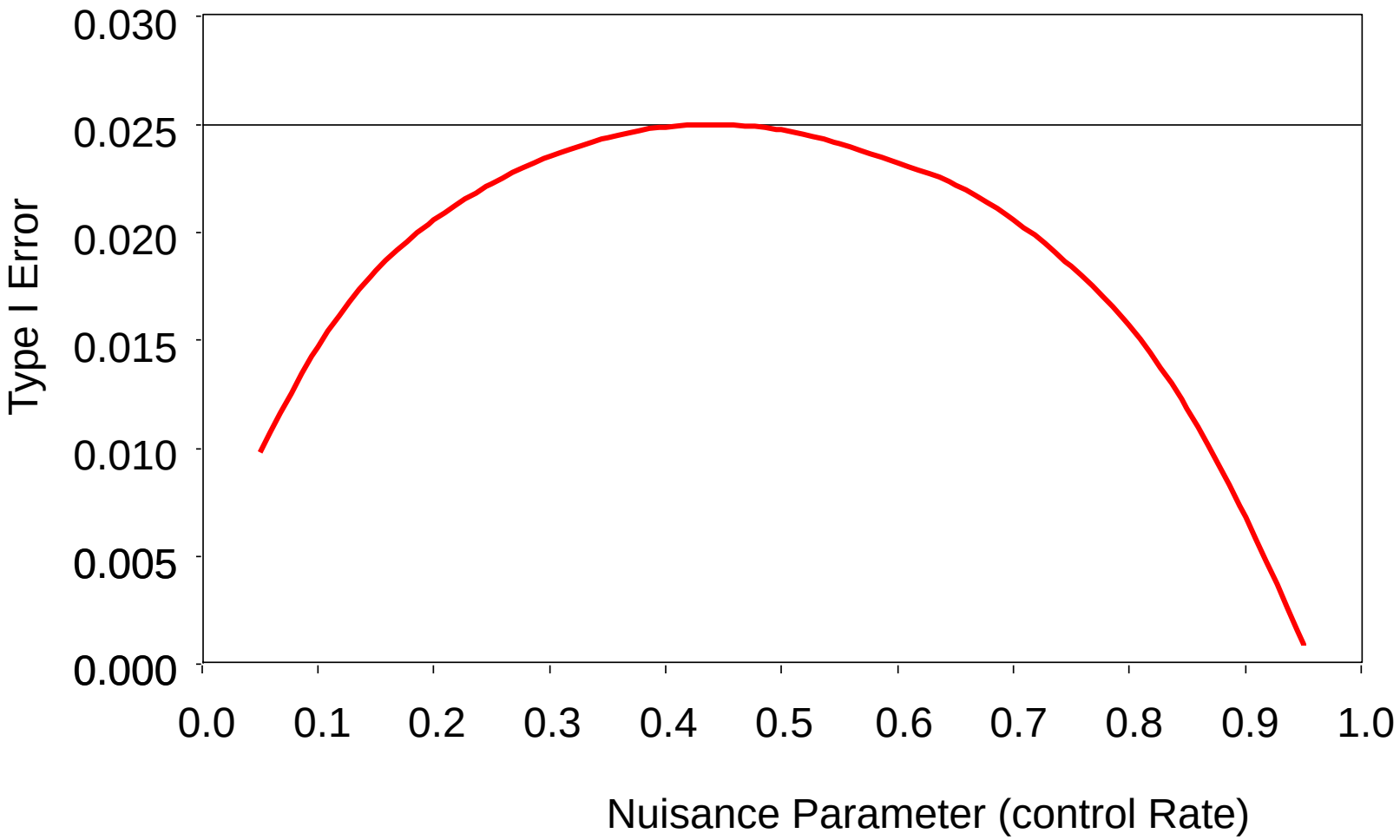
Bayesian AD – Thall & Wathen(EJC, 2007) N=200
Critical Value to Control One-Sided Type-I Error
10^6 Simulations / control rate



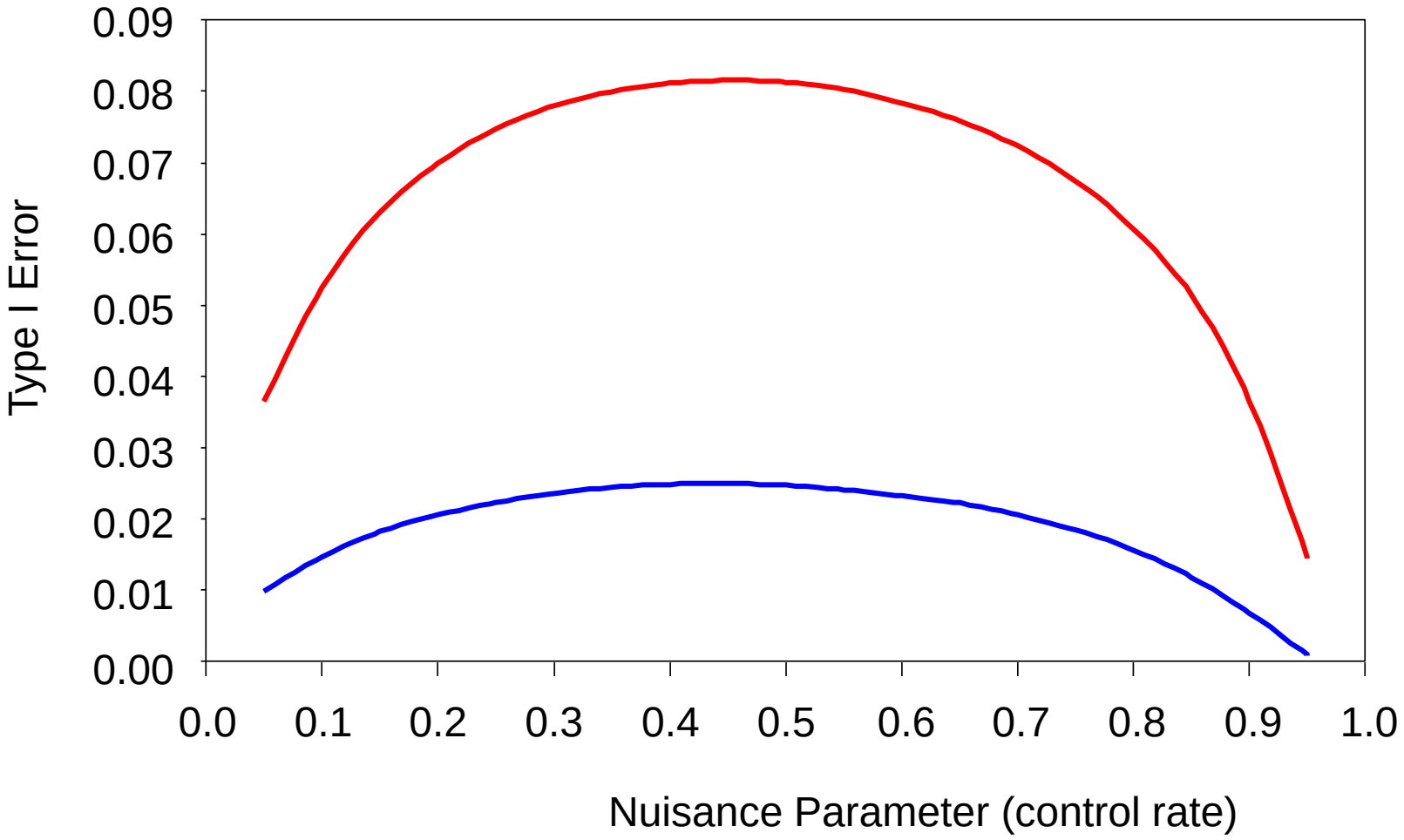
Bayesian AD – Thall & Wathen(EJC, 2007) N=200
Critical Value to Control One-Sided Type-I Error
10^6 Simulations / control rate



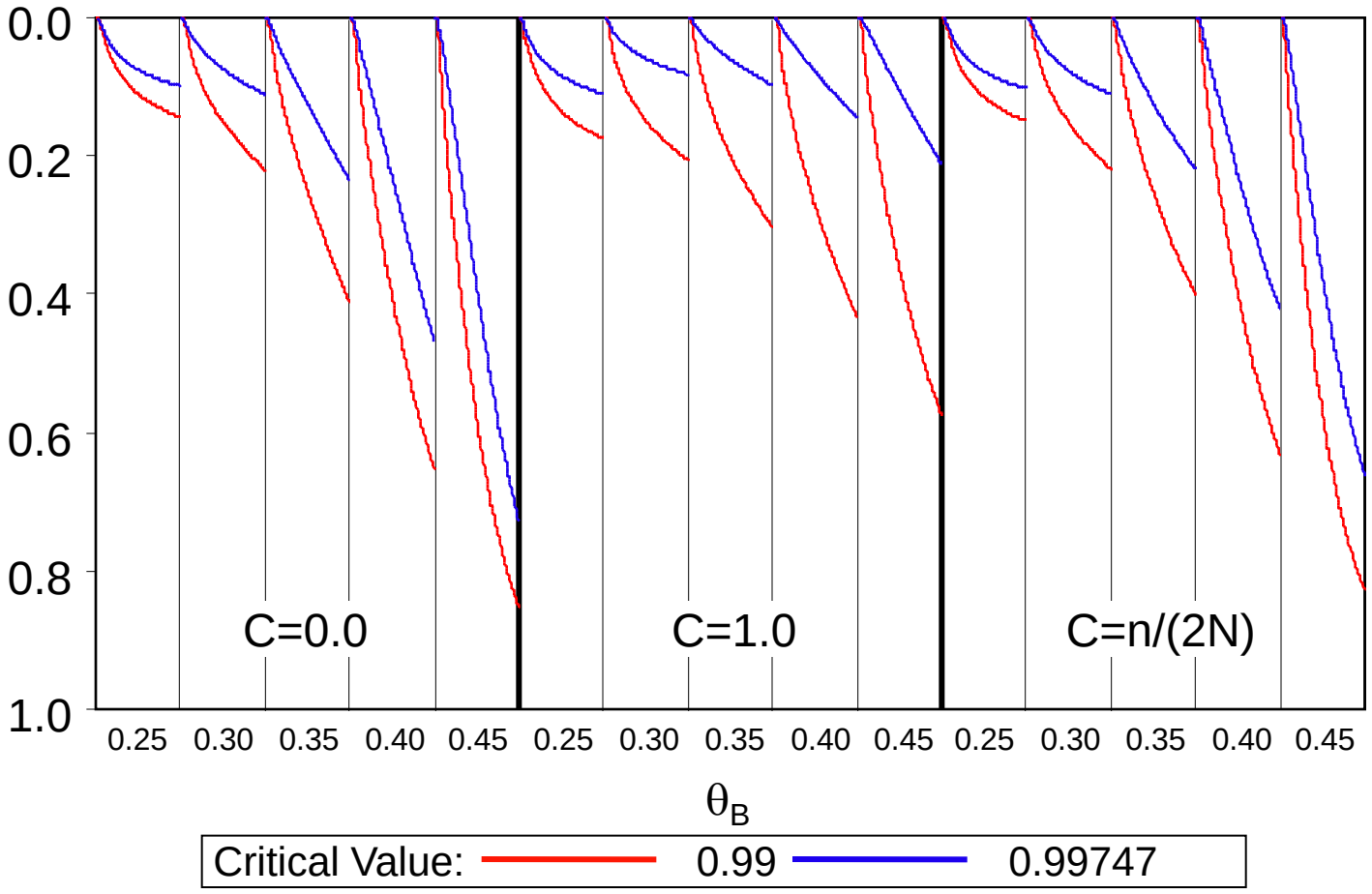
Bayesian AD – Thall & Wathen(EJC, 2007) N=200
Type-I Error Based on $P(\pi_A<\pi_B|Data)>.99747$
10^6 Simulations / control rate



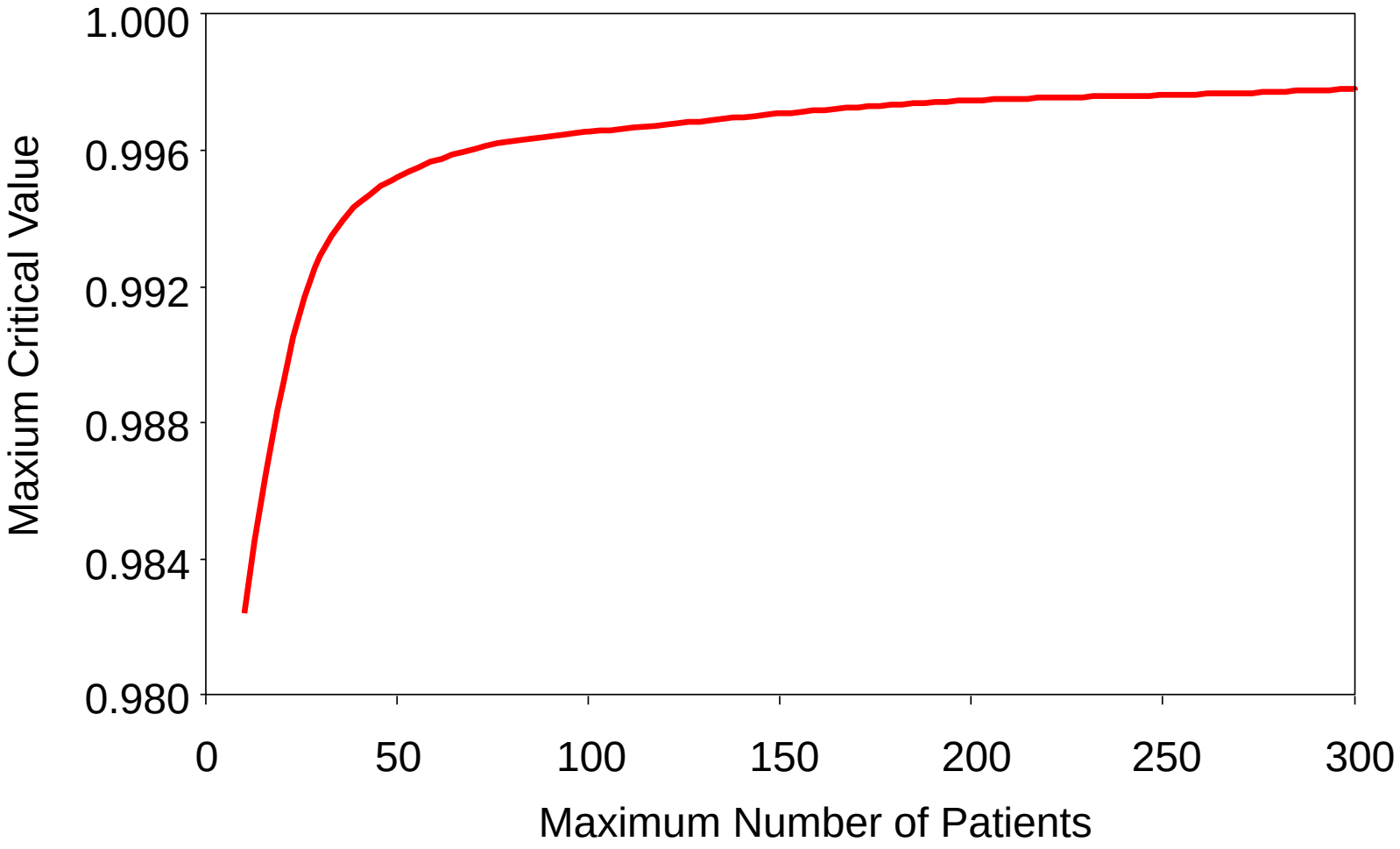
Bayesian AD – Thall & Wathen(EJC, 2007) N=200 Comparison of Type-I Error Based on T&W Criterion & Adjusted



Bayesian AD – Thall & Wathen(EJC, 2007) N=200
Probability of Choosing B as a Function of Patient Accrual
10^5 Simulations



Bayesian AD – Thall & Wathen(EJC, 2007) Maximum Critical Value vs Maximum Patient Number >10^6 Simulations / control rate



Crisiticism of This Approach

- Korn and Freidlin (J Clin Oncol, 2011)
- Their simulations “show”:
  - Thall & Wathen AD inferior to1:1 randomisation in terms of information, benefits to patients in trial

- True
- I agree with Don Berry (J Clin Ocol 2011) that the greatest benefits are likely to accrue for trials with more than 2 arms
- Rather as in the case of T=1 in the group sequential case greater complexity gives more scope for Bayesian designs

Setting Up Simulations

- How do we design a simulation experiment ?
- What are the factors ?
- Can we use a fractional factorial

ASTIN Trial – Acute Stroke: Dose Effect Curve (Grieve and Krams, Clinical Trials, 2005) Factors of the Simulation Experiment

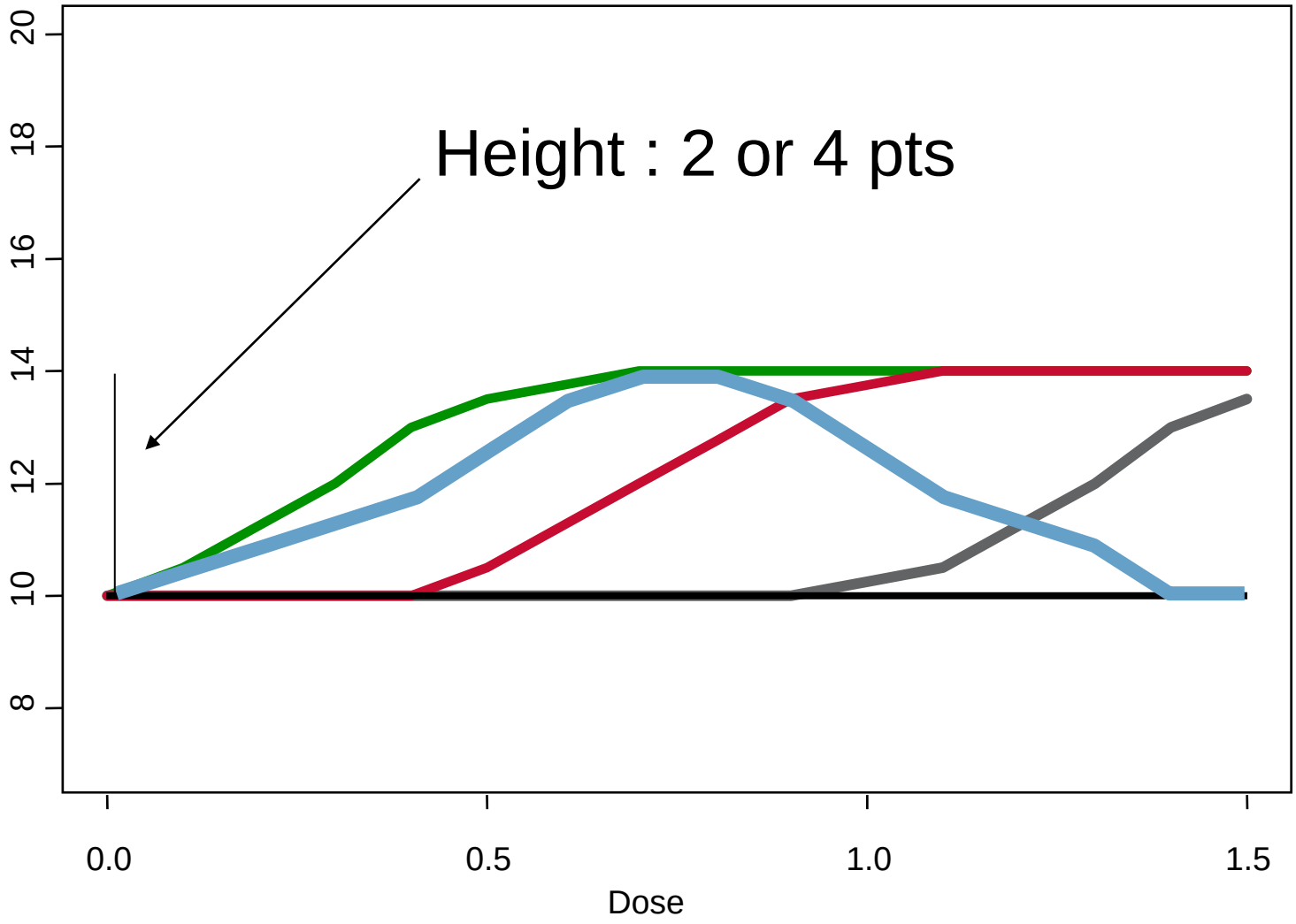
- Termination Rule
  - Decision theoretic
  - Posterior Probabilities
- Prior Distribution
  - Uninformative (flat) prior
  - Informative prior
- Variance of Prior
  - Constant
  - Less variable at low end of the dose range
- Level of Smoothing
  - High smoothing
  - Low Smoothing
- Allocation Criteria
  - Var(ED95) x Var(f(z*))
  - Var(f(z*))
  - Var(ED95)
  - Determinant of Covar ((ED95) and Var(f(z*))
- Randomisation Method
  - Probability of allocation proportional to expected utility
  - Allocate to maximum utility
  - Probability uniform over doses st $0.9f(z_0) < E[f(z)|Y] < 1.10f(z_0)$
  - Probability uniform over doses st $0.9f(z_0) < E[f(z)|Y]$
- $2^4 \times 4^2$ experiment 1/4 replicate

ASTIN Trial – Acute Stroke: Dose Effect Curve (Grieve and Krams, Clinical Trials, 2005) Aliasing Structure

Contrast	Effects	Contrast	Effects
1	A, BCDG, BEFH	19	BF, AEH
2	B, ACDG, AEFH	20	BG, ACD
3	C, ABDG	21	BH, AEF
4	D, ABCG	22	CD, ABG, EFGH
5	E, ABFH	19	BF, AEH
6	F, ABEH	20	BG, ACD
7	G, ABCD	21	BH, AEF
8	H, ABEF	22	CD, ABG, EFGH
9	AB, CDG, EFH	23	CE, DFGH
10	AC, BDG	24	CF, DEGH
11	AD, BCG	25	CG, ABD, DEFH
12	AE, BFH	26	CH, DEFG
13	AF, BEH	27	DE, CFGH
14	AG, BCD	28	DF, CEGH
15	AH, BEF	29	DG, ABC, CEFH
16	BC, ADG	30	DH, CEFG
17	BD, ACG	31	EF, ABH, CDGH
18	BE, AFH	32	EG, CDFH

ASTIN Trial – Acute Stroke: Dose Effect Curve (Grieve and Krams, Clinical Trials, 2005)

Dose Response Curves



Simulation Experiment

- Simulated
  - 20 replicates
  - 64 experiments
  - 9 curves
    - 11520 individual studies
- Run-on a network of PCs/Workstations
- Having established “optimal” settings
  - 1000 simulations were performed to establish operating characteristics

Accuracy of Simulations
Posch, Mauerer and Bretz (SIM, 2011)

- Studied adaptive design with treatment selection at an interim and sample size re-estimation
- Control FWER (familywise error rate) in a strong sense – under all possible configurations of hypotheses

- Conclude: That you have to be careful with the assumptions behind the simulations.

- Intriguing point: the choice of seed has an impact on the estimated type I error

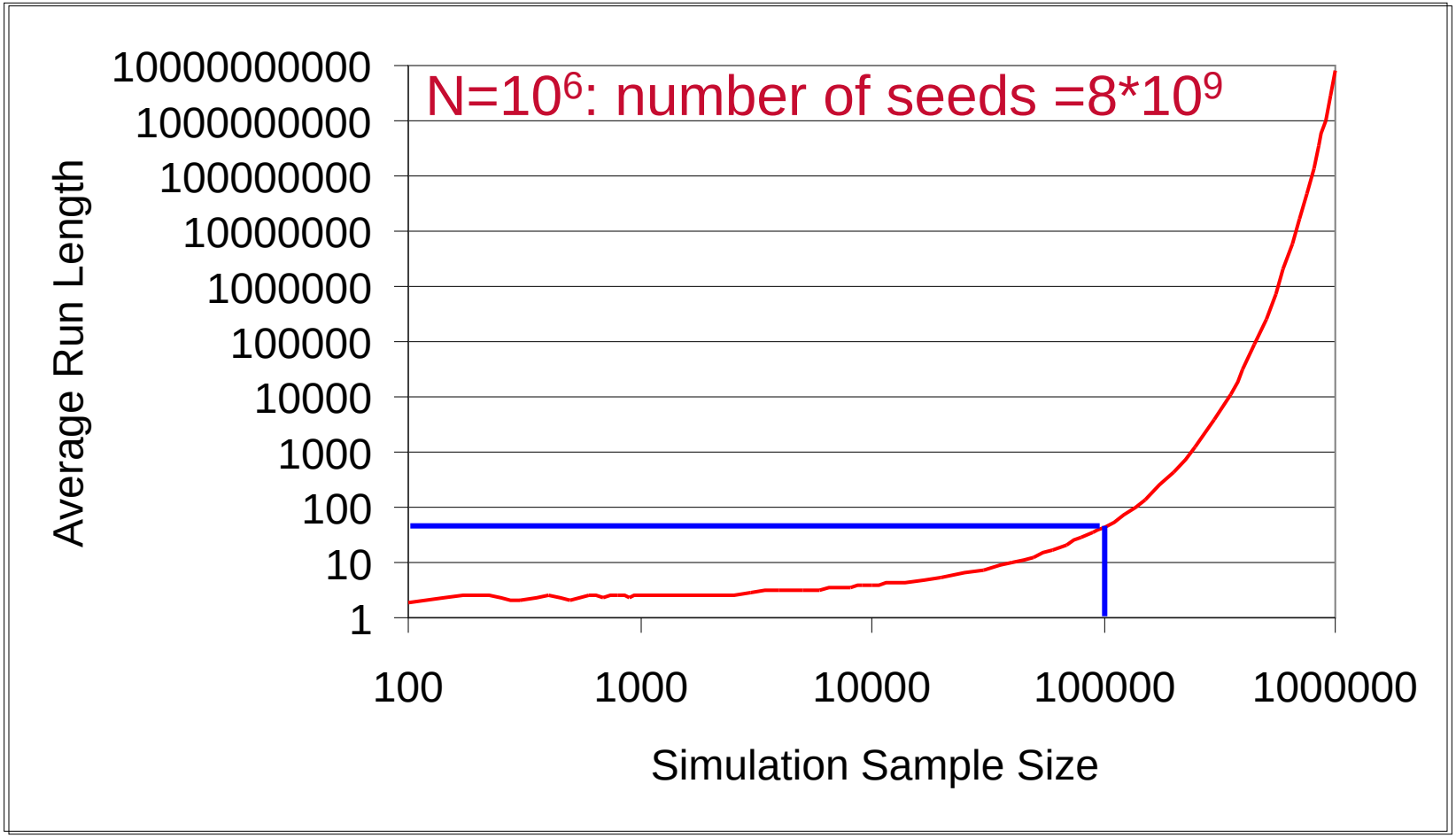
Accuracy of Simulations
Posch, Mauerer and Bretz (SIM, 2011)

- If it is important to be able to differentiate between 0.025 and 0.026 then we should power our simulation experiment for it

- A sample of 10^4 has only 10% power to detect H_A=0.026 vs H_0=0.025, 10^5: 50%

- 80% power requires n=194,000 – search 380 seeds
- 90% power requires n=260,000 – search 1600 seeds

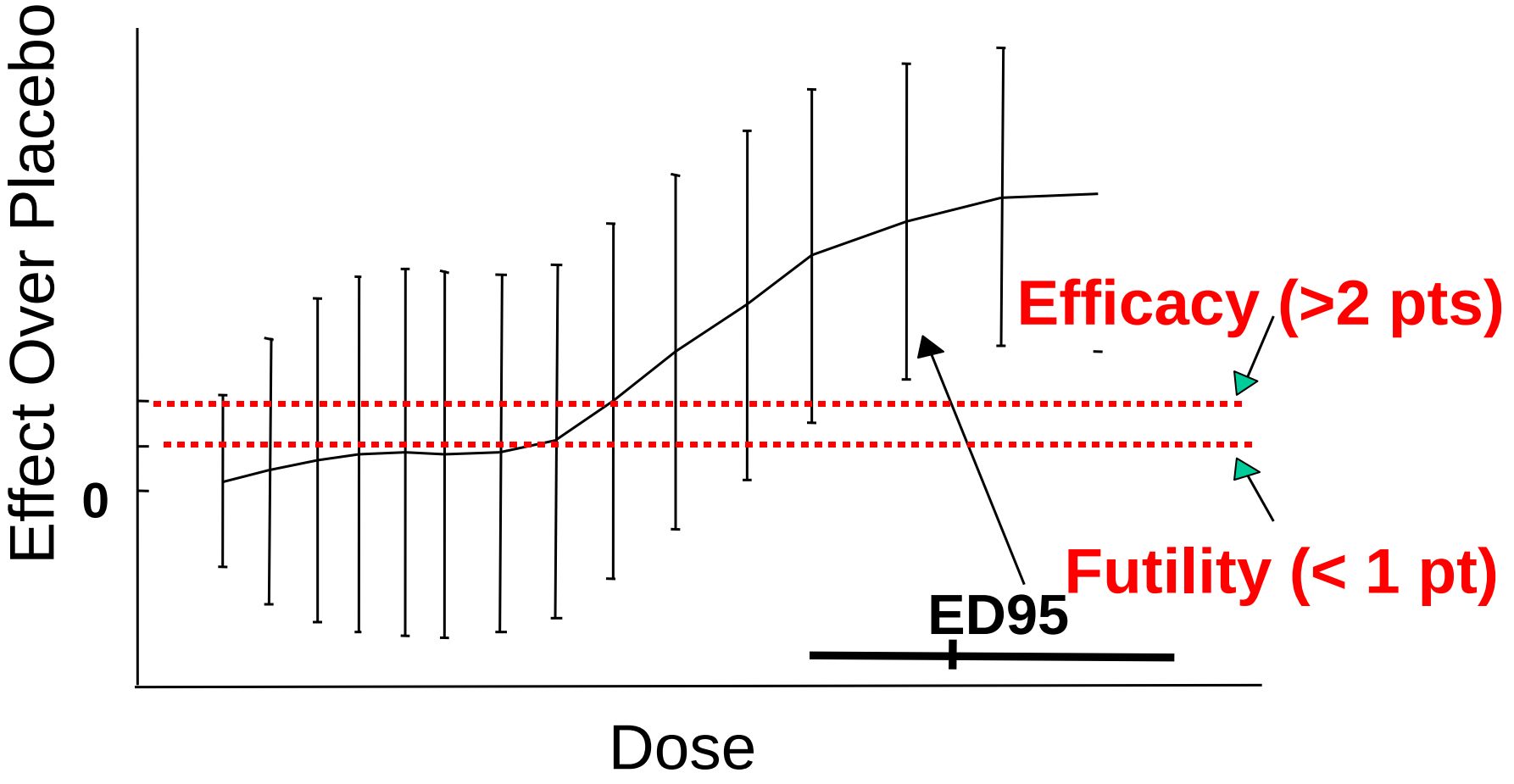
Average Run Length to find a "Good Seed"



Conclusions
Determining Decision Criteria

- Appropriate approach:
  - Choose decision rule based on clinical or commercial criteria.
  - Investigate operating characteristics
  - If they are unacceptable e.g., type I error > 20% then look to change them
  - BUT don’t strive to get exact control
- Two examples

ASTIN Trial – Acute Stroke: Dose Effect Curve (Grieve and Krams, Clinical Trials, 2005)



POC Study in Neuropathic Pain
Smith et al (Pharmaceutical Statistics, 2006)