Frequentist Equivalence and Bayesian Advantages

Evaluating a Fully Bayesian Adaptive Design Against Its Hybrid Counterpart

Jose Manuel Lema Martinez[1,2] Kevin Wilson[2] Nina Wilson[3]

[1]NDORMS, University of Oxford [2]School of Mathematics, Statistics and Physics, Newcastle University
[3]Population Health Sciences Institute, Newcastle University

Bayes Pharma 2026





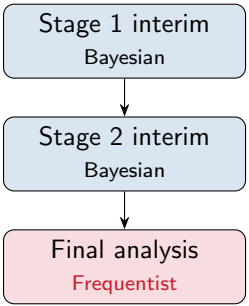
Most Bayesian adaptive trials stop being Bayesian at the end

The hybrid design

- Interim: posterior & predictive probabilities decide stop / continue
- Final: frequentist test, p-value
- Reason: primary inference expressed as type I error & power

The cost: inferential coherence

- Beliefs updated by Bayes' theorem ... then discarded at the decision that matters



Jan 2026 FDA draft guidance: Bayesian primary inference is acceptable if operating characteristics are shown prospectively by simulation

The question

What happens to a trial's operating characteristics
when its hybrid design is made **fully Bayesian**?

1. **Frequentist properties:** can a fully Bayesian design keep the hybrid's type I error and power?
2. **Bayesian properties:** what do assurance and Bayesian type I error tell us that power and type I error do not?
3. **Design choices:** how do the final threshold and the allocation of patients across stages affect performance?

Case study: the WIRE platform trial

- Phase II, multi-arm, non-randomised platform
- Renal cell carcinoma, window of opportunity before surgery
- Up to 3 stages per arm, binary endpoint
- Two arm types, analysed separately

	Monotherapy	Combination
H_0: π_0	0.20	0.30
H_1: π_1	0.50	0.60
Max N per arm	12	20
One-sided α	0.10	0.05

Hybrid vs fully Bayesian: only the final rule changes

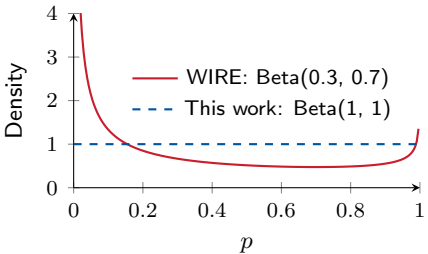
	Hybrid	**Fully Bayesian**
Interim efficacy	$P(p_k > \pi_0 \mid \text{data}) > 0.98$	*identical*
Interim futility	$PP_k < 0.02$, target = final test	$PP_k < 0.02$, target = final posterior rule
Final analysis	Exact binomial test, $p \leq \alpha$	$P(p_k > \pi_0 \mid \text{data}) > \gamma_S$
Analysis prior	Beta(1, 1) for all posterior and predictive calculations	
Stages, $n_{s,k}$	*identical*	

- $\gamma_S \in \{0.90,\ 0.95\}$ for both arm types
- PP_k: predictive probability of meeting the final success rule if all remaining patients are recruited

Priors: two jobs

Analysis prior

- Beta(1,1) for every arm, both designs
- Replaces WIRE's U-shaped Beta(0.3,0.7)



Design prior

Beta with mean π_0 or π_1, variance σ^2

(a) Conditional on a hypothesis

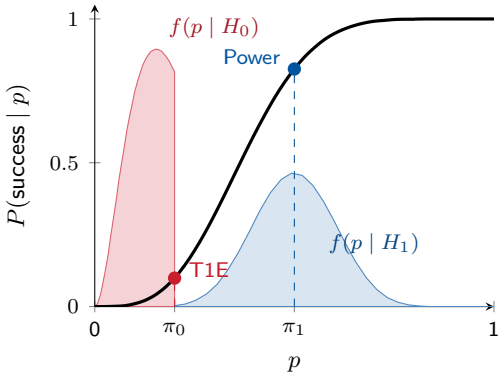
- H_0: truncated to $[0, \pi_0] \rightarrow$ Bayesian T1E
- H_1: truncated to $(\pi_0, 1] \rightarrow$ assurance

(b) Mixture: effect genuinely unknown

$$f(p_k) = w\, f(p_k \mid H_0) + (1 - w)\, f(p_k \mid H_1)$$

- $w = P(H_0)$: probability the drug does not work
- $\rightarrow$ unconditional assurance, $w \in \{0.1, \ldots, 0.9\}$

Operating characteristics: points vs averages



Frequentist: read the curve at a point

$$\text{T1E} = P(\text{success} \mid \pi_0) \qquad \text{Power} = P(\text{success} \mid \pi_1)$$

Bayesian: average the curve over a design prior

$$\text{BT1E} = \int_0^{\pi_0} P(\text{success} \mid p)\, f(p \mid H_0)\, dp$$

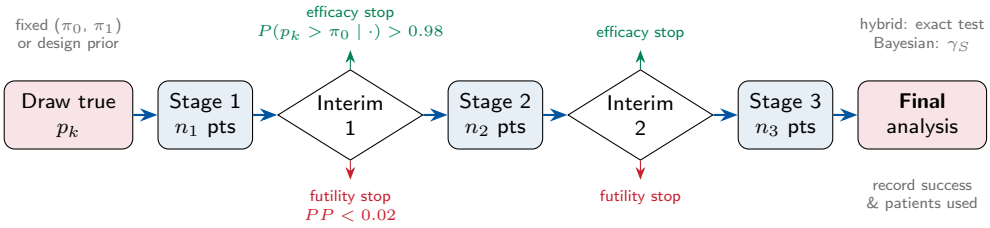
$$\text{Assurance} = \int_{\pi_0}^{1} P(\text{success} \mid p)\, f(p \mid H_1)\, dp$$

Mixture: unconditional assurance (UA)

$$\text{UA} = w\,\text{BT1E} + (1 - w)\,\text{Assurance}$$

As $\sigma^2 \to 0$, Bayesian $\to$ frequentist.

Simulation study



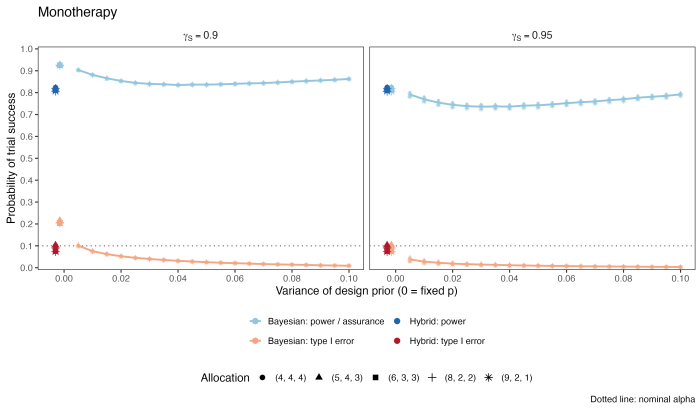
Stage allocations (n_1, n_2, n_3)

Mono ($N = 12$)	Combo ($N = 20$)
(4, 4, 4)	(7, 7, 6)
(5, 4, 3)	(8, 6, 6)
(6, 3, 3)	(10, 5, 5)
(8, 2, 2)	(12, 4, 4)
(9, 2, 1)	(14, 3, 3)

balanced $\rightarrow$ front-loaded

- $J = 100{,}000$ trials per configuration
- Both designs applied to the same simulated patients
- Final threshold $\gamma_S \in \{0.90,\ 0.95\}$
- Design-prior variance $\sigma^2 \in [0.005,\ 0.1]$ (ESS and mixture: $[10^{-4},\ 0.1]$)
- Mixture weight $w \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$

Result 1: equivalence in monotherapy arms

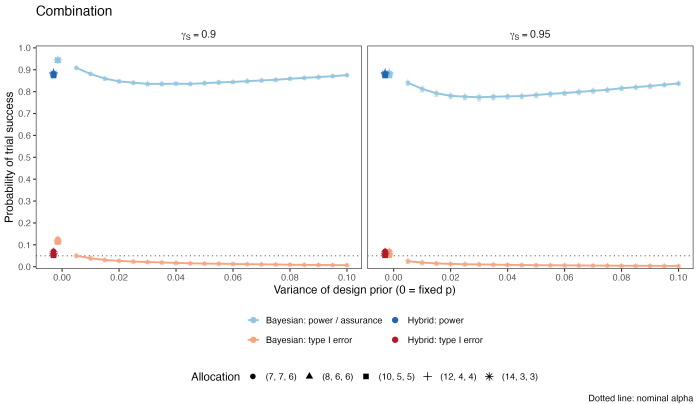


At variance 0 (fixed p)

	T1E	Power
Hybrid	0.07–0.10	0.81–0.83
$\gamma_S = 0.90$	0.20–0.21	0.93
$\gamma_S = 0.95$	0.07–0.10	0.81–0.83

Calibrated correctly, going fully Bayesian costs nothing in frequentist terms

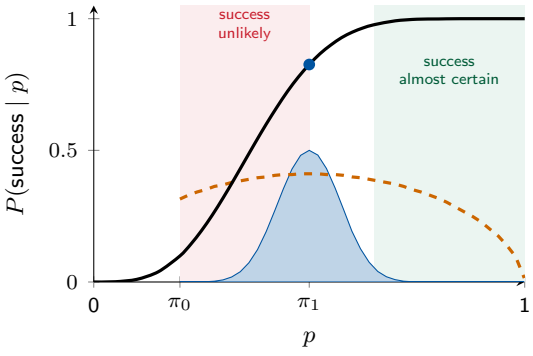
Result 1 (cont.): same pattern for combination arms



At variance 0 (fixed p)

	T1E	Power
Hybrid	0.05–0.07	0.87–0.89
$\gamma_S = 0.90$	0.11–0.13	0.94–0.95
$\gamma_S = 0.95$	0.05–0.07	0.87–0.89

Result 2: what the Bayesian metrics add



Assurance: power averaged over $f(p \mid H_1)$

- Wider prior: mass moves into low and high regions
- Losses outweigh gains → dips, partly recovers, *stays below power*

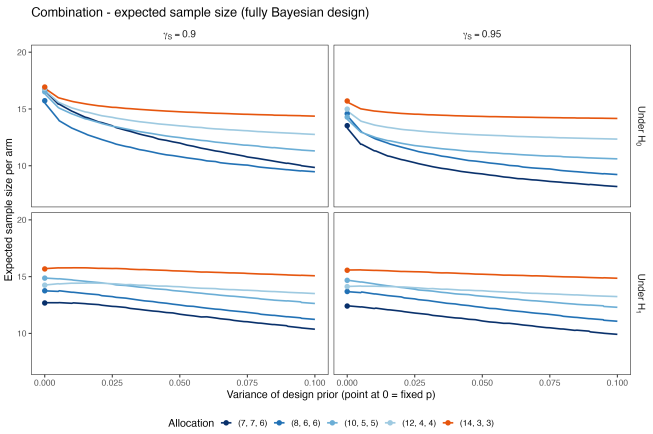
γ_S=0.95: mono 0.79→0.74→0.79 (power 0.82)
combo 0.84→0.78→0.84 (power 0.88)

Bayesian T1E: T1E averaged over $f(p \mid H_0)$

- Prior sits below π_0, where success is rare
- Point T1E is the *worst case*

γ_S=0.95: mono 0.04, combo 0.03 → < 0.01

Result 3: balanced allocation lowers expected sample size



Combination arms

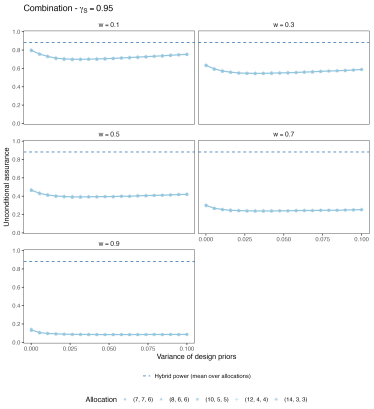
- Under H_1: strict ordering, (7,7,6) lowest, (14,3,3) highest
- Under H_0: (7,7,6) lowest at $\gamma_S = 0.95$; (8,6,6) at $\gamma_S = 0.90$
- ESS falls with σ^2

Monotherapy (backup)

- Under H_1: (4,4,4), (5,4,3) lowest
- Under H_0: ranking depends on γ_S

Same total N, same rules: front-loading is consistently the most expensive

Result 4: unconditional assurance when the effect is unknown



- Uncertainty about **whether** the drug works (w) matters far more than uncertainty about the **size** of its effect (σ^2):
  combo, $\gamma_S = 0.95$: $0.80 \to 0.14$ across w; at most 0.10 along any curve
- A probability of success that doesn't assume efficacy:
  $w = 0.5 \Rightarrow$ roughly a 40% chance the trial succeeds
- Essentially unaffected by allocation
- Same dip-then-flatten shape as assurance

Monotherapy: same behaviour (backup)

A probability of success that does not assume the drug works

Implications and limitations

What we found

- **Frequentist equivalence:** with a well-calibrated γ_S, the fully Bayesian design has the same type I error, power and ESS as the hybrid design
- **Bayesian advantage:** assurance and Bayesian type I error do not assume a fixed response rate; they average over uncertainty about the treatment effect
- **Allocation:** more balanced allocations need fewer patients when the treatment works, with a similar probability of success; front-loading is worst

Alignment with FDA draft guidance (2026)

- Calibrated to type I error (§1), with Bayesian OCs reported on top (§2)
- Transparent priors: flat analysis prior; design priors centred on the response rates assumed in the WIRE protocol

Limitations

- Setting modelled on WIRE: binary endpoint, small N, arms analysed independently against π_0
- Flat analysis prior makes equivalence achievable; informative priors or borrowing would change decisions

Thank you!

Thank you for your time and consideration!

jose.lemamartinez@ndorms.ox.ac.uk

Frequentist Equivalence and Bayesian Advantages: Evaluating a Fully Bayesian Adaptive Design Against Its Hybrid Counterpart. Lema, Wilson & Wilson (under review)

Backup: with binary data, every final rule is a responder cut-off

Final rule	Mono ($N = 12$, $\pi_0 = 0.2$)	Combo ($N = 20$, $\pi_0 = 0.3$)
Exact binomial, $P(X \geq x) \leq \alpha$	$x \geq 5$	$x \geq 10$
Posterior, $\gamma_S = 0.90$	$x \geq 4$	$x \geq 9$
Posterior, $\gamma_S = 0.95$	$x \geq 5$	$x \geq 10$

Example: 4 responders out of 12

- p-value $= 0.205 \rightarrow$ not significant at $\alpha = 0.10$
- $P(p > 0.2 \mid 4/12) = 0.901 \rightarrow$ passes $\gamma_S = 0.90$
- Posterior $\neq 1-$ p-value

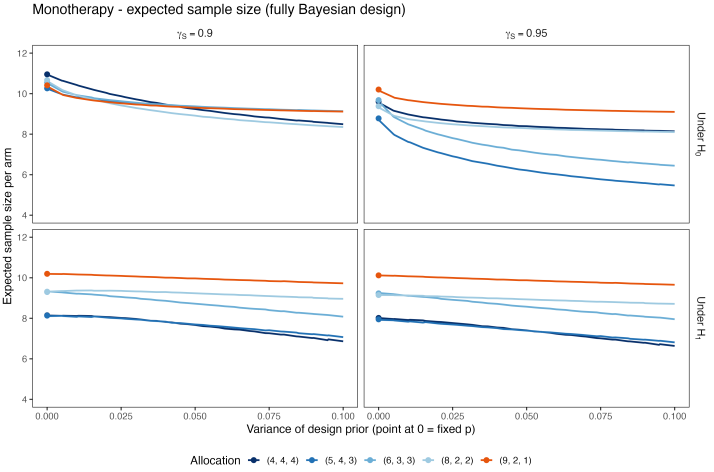
Exact identity, Beta(1,1) prior

$$P(p > \pi_0 \mid x, N) = P\big(\mathsf{Bin}(N+1, \pi_0) \leq x\big)$$

A posterior rule is an exact test with *one extra responder* $\Rightarrow$ more liberal at $\gamma_S = 1 - \alpha$

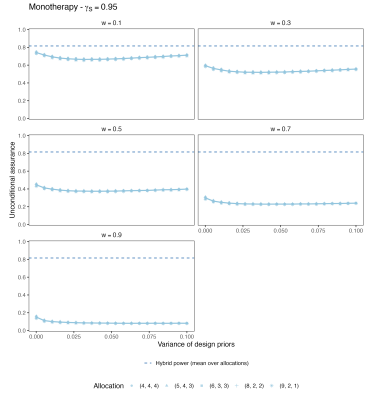
Calibrate γ_S by matching rejection regions, not by setting $\gamma_S = 1 - \alpha$

Backup: monotherapy expected sample size

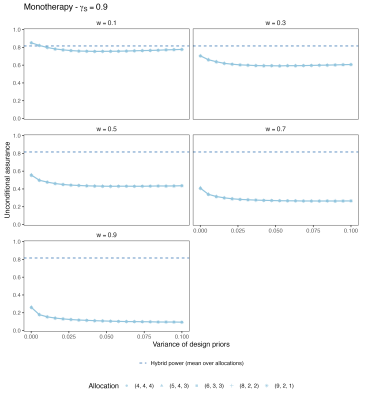


Under H_0, $\gamma_S = 0.90$: (4,4,4) highest at $\sigma^2 = 0$ (no stage-1 futility stop possible with cut-off 4);
$\gamma_S = 0.95$: (5,4,3), (6,3,3) lowest.

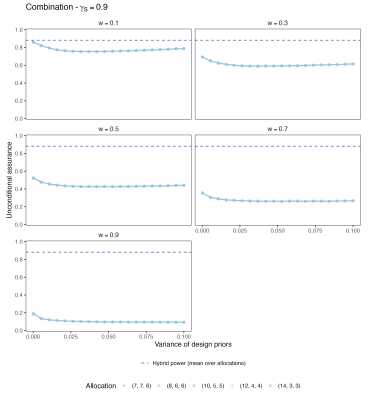
Backup: unconditional assurance, other panels



Mono, $\gamma_S = 0.95$



Mono, $\gamma_S = 0.90$



Combo, $\gamma_S = 0.90$

Backup: predictive probability for futility

Posterior $p_k \mid \text{data} \sim \text{Beta}(a, b)$, $a = 1 + S_{s,k}$, $b = 1 + F_{s,k}$; $m = N_k - n_{s,k}$ patients remaining.

$$P(Y = j \mid \text{data}) = \binom{m}{j} \frac{B(a + j,\, b + m - j)}{B(a, b)}$$

$$PP_k = \sum_{j=0}^{m} P(Y = j \mid \text{data})\, \mathbb{I}[\text{final success with } S_{s,k} + j \text{ responders}]$$

Stop for futility if $PP_k < 0.02$. Final success = exact test (hybrid) or posterior $> \gamma_S$ (Bayesian).

Backup: frequentist operating characteristics by allocation

		Hybrid = Bayesian $\gamma_S = 0.95$		Bayesian $\gamma_S = 0.90$	
Arm	Allocation	T1E	Power	T1E	Power
Mono	(4,4,4)	0.099	0.826	0.209	0.928
	(5,4,3)	0.102	0.820	0.214	0.927
	(6,3,3)	0.076	0.808	0.203	0.925
	(8,2,2)	0.091	0.821	0.205	0.927
	(9,2,1)	0.073	0.806	0.203	0.925
Combo	(7,7,6)	0.072	0.881	0.126	0.946
	(8,6,6)	0.061	0.881	0.116	0.941
	(10,5,5)	0.052	0.873	0.114	0.943
	(12,4,4)	0.068	0.886	0.121	0.945
	(14,3,3)	0.058	0.881	0.116	0.944