A Bayesian Precision-Medicine Decision Framework for Pursuing Biologically Plausible Predictive Biomarkers in Early Clinical Development

Mathias Cardner[1], Benjamin Georgi[1], Bastian Angermann[1], Chris Chamberlain[2], Adam Platt[2], Daniel Muthas[1]

Translational Science and Experimental Medicine, Research and Early Development, Respiratory & Immunology, BioPharmaceuticals R&D, AstraZeneca, [1]Gothenburg, Sweden, [2]Cambridge, UK

All authors are/were employees of AstraZeneca, and may own stock options

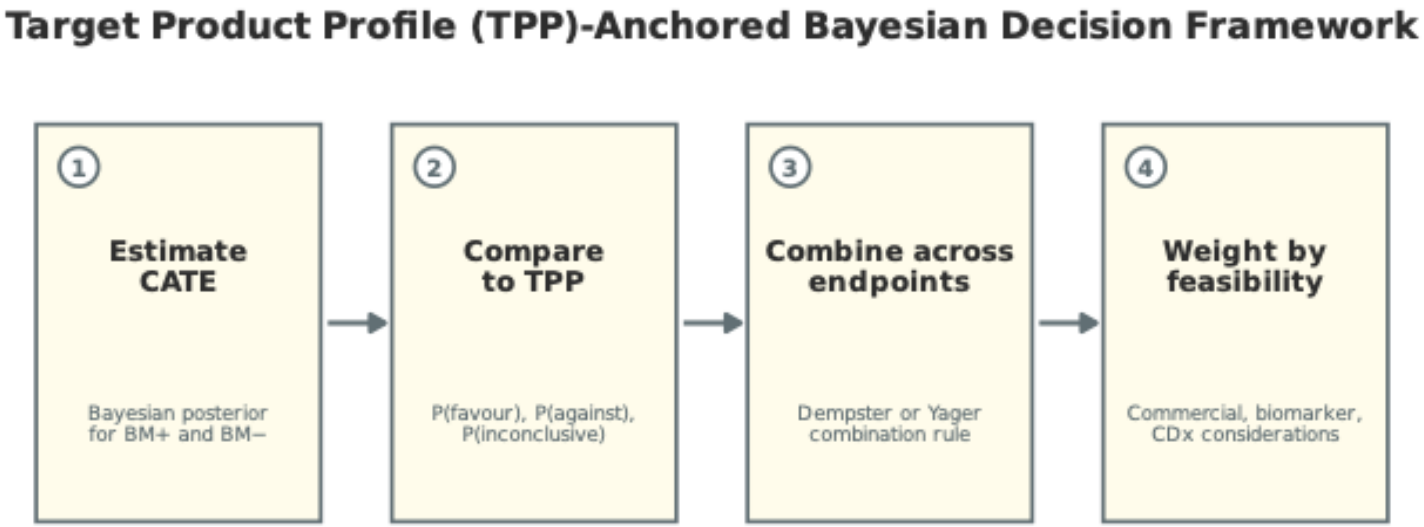
This presentation is based entirely on synthetic data and simulations

The European Medicines Agency acknowledges that:

"currently available [statistical *subgroup-efficacy*] tests lack power (sensitivity) to detect […] effects that are of potential clinical importance […] Biological plausibility and the ability to find replication are key elements to evaluate the credibility of a subgroup finding."

Source: https://www.ema.europa.eu/en/investigation-subgroups-confirmatory-clinical-trials-scientific-guideline

We propose a Bayesian framework for evaluating evidence of precision-medicine propositions in early clinical trials



A single (S)-learner can quantify the treatment effects in biomarker-defined subgroups

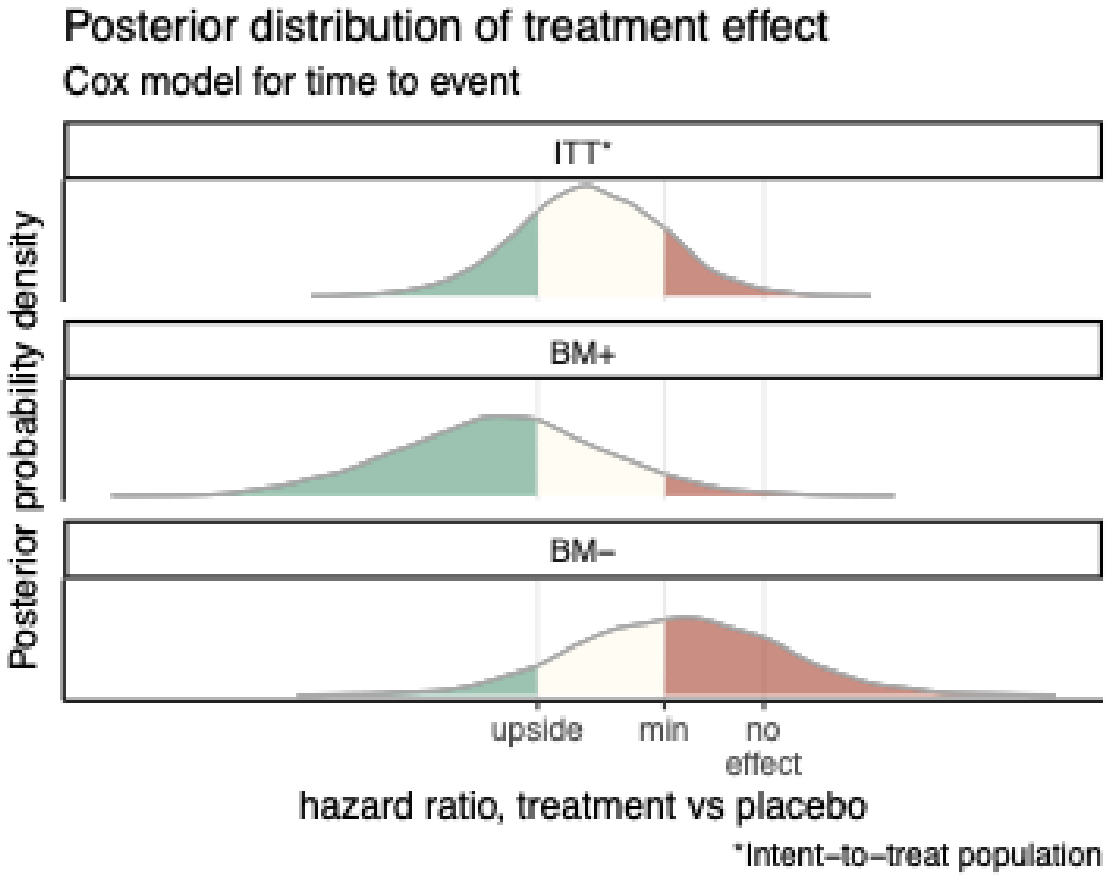
- The clinical outcome Y is regressed on binary indicator variables of treatment T (active=1 vs placebo=0), biomarker B (BM+ vs BM-), and their interaction. Covariates are modelled as main effects:

$$\text{link}(\mathbb{E}Y) = \alpha + \tau T + \beta B + \delta TB + \boldsymbol{\gamma}^T \text{covariates}$$

- The BM×Treatment interaction-term coefficient δ can be tested classically (frequentist)
- The conditional average treatment effect (**CATE**) aggregated per subgroup is:
  - in biomarker negative: $\mathbb{E}[Y(1) - Y(0) \mid B = \text{BM}-] = \tau$
  - in biomarker positive: $\mathbb{E}[Y(1) - Y(0) \mid B = \text{BM}+] = \tau + \delta$
- We propose relating the **CATE posterior to the target-product-profile (TPP) cases** for efficacy (upside, base, min)

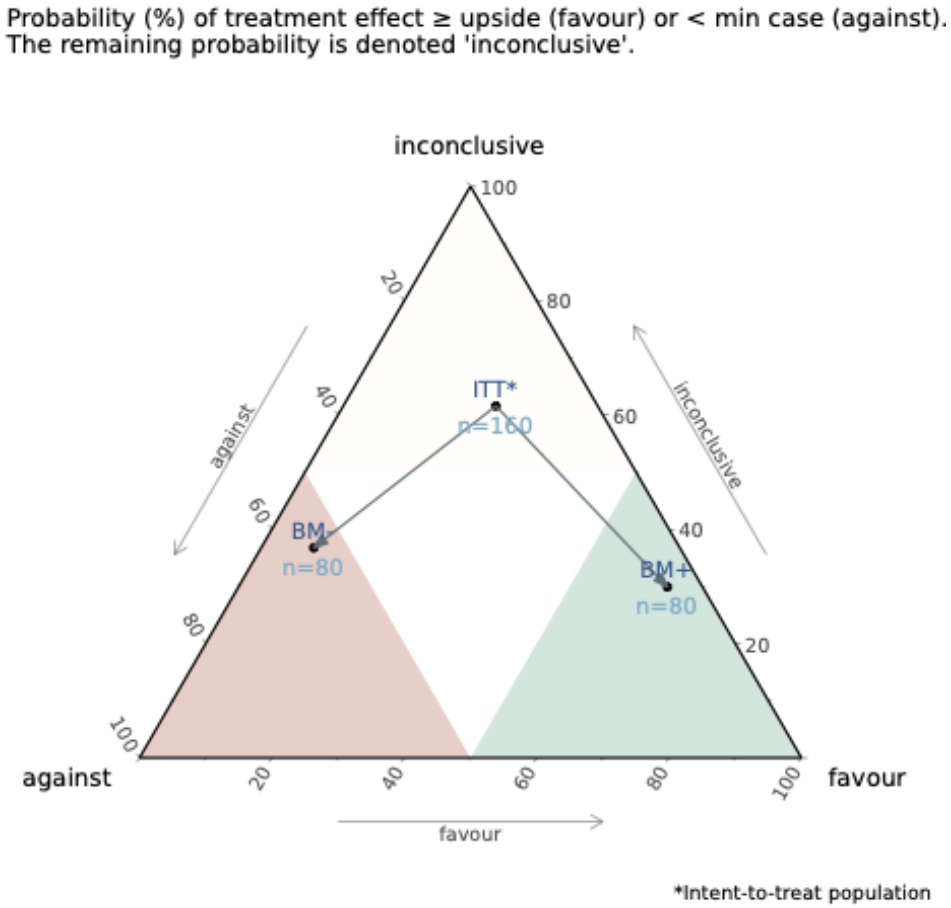
The CATE posterior distribution encodes rich information on the evidence for efficacy in subgroups

- *Single simulated trial: N=160, HR(BM+)=0.65, HR(BM−)=0.80*
- BM+ posterior: **65% P(favour)**, 5% P(against)
- Frequentist interaction: p=0.207 — “non-significant”



A ternary plot encodes posterior evidence (favour, against, inconclusive) in a single point

- The BM+ point is in green, meaning P(favour) ≥ 50%
- The ITT is inconclusive, meaning P(favour) + P(against) < 50%
- The BM− points is in red, meaning P(against) ≥ 50%
- Visual summary of the evidence landscape
- *Same simulated trial as on previous slide*

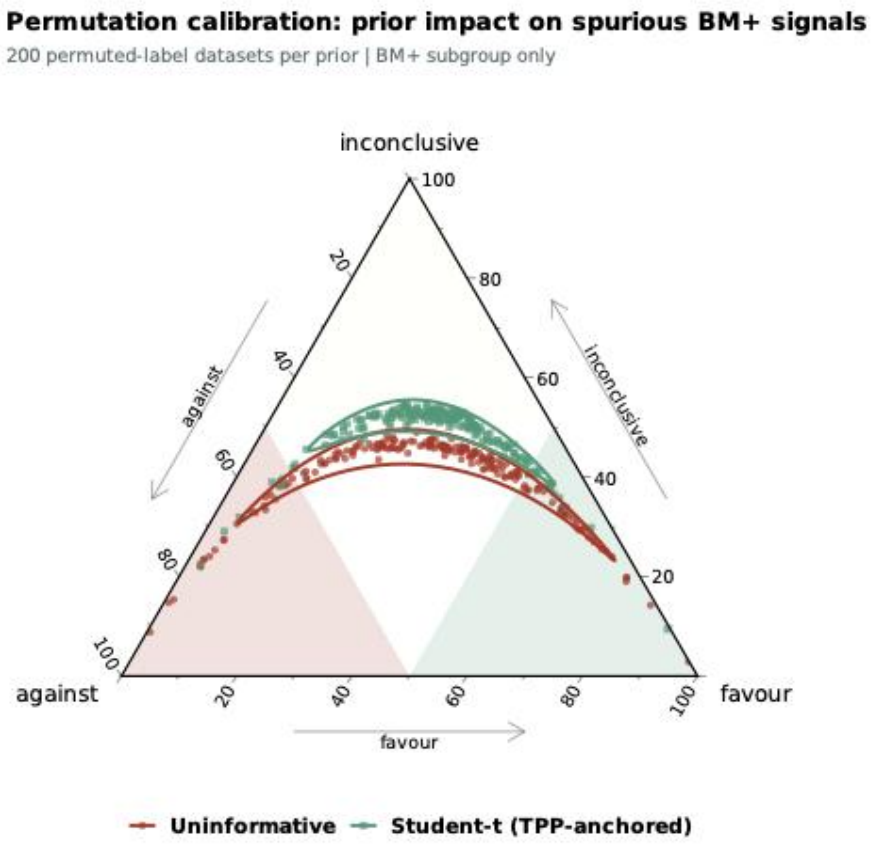


Simulation study to characterise the Bayesian framework's operating properties

- Fully synthetic Phase-2 trials: N = 100–260, 1:1 randomisation
- Binary predictive biomarker (BM+ prevalence 20–70%)
- Two endpoints: time-to-event (Cox PH) and change-from-baseline (Gaussian)
- Bayesian S-learner: Treatment × BM interaction
- 200–500 simulations per scenario
- Sensitivity vs frequentist testing: HR(BM+), N, prevalence
- Specificity under the null: Prior family, permutation
- Multi-endpoint combination: Dempster vs Yager rules
- Prior calibration: SD sweep, df sweep
- Multiple biomarker panels: p = 3, 5, 10 candidates

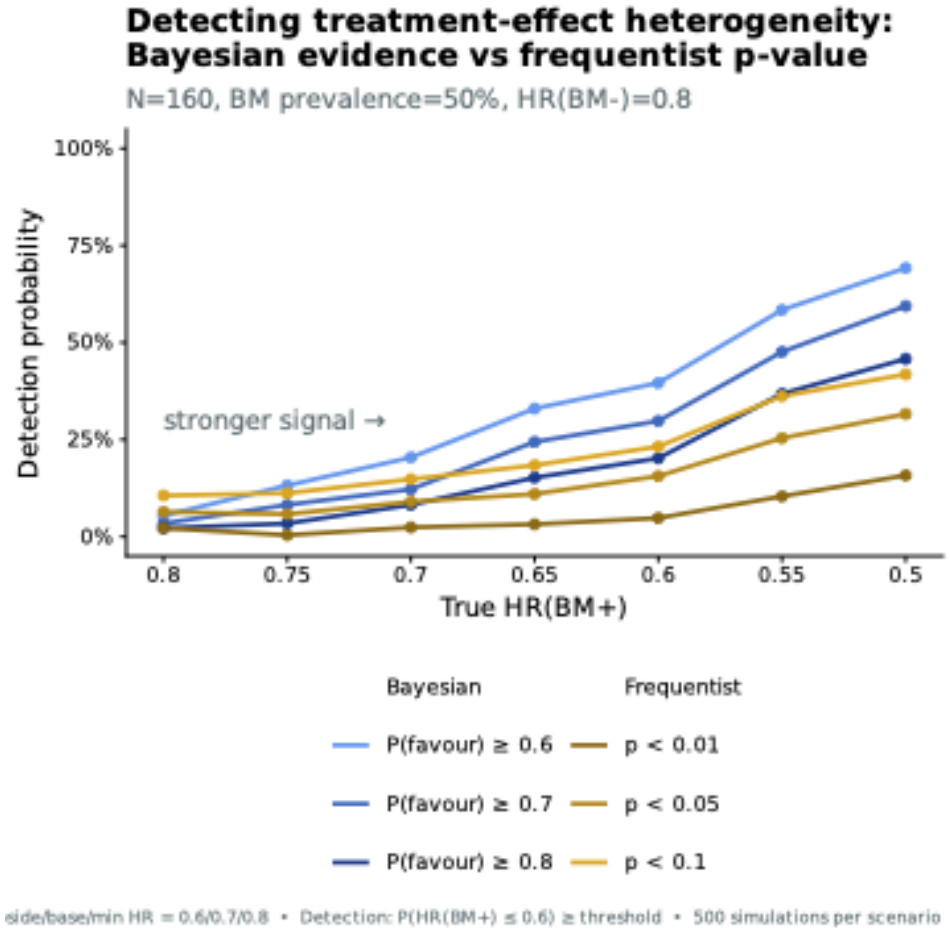
TPP-anchored priors shrink spurious heterogeneity, preserve true signals

- **Red** (uninformative): wider scatter into favour region
- **Green** (Student-t): points cluster near inconclusive
- Ellipses show 80% coverage
- Prior acts as built-in multiplicity control
- *200 permuted-label datasets per prior; BM+ subgroup*



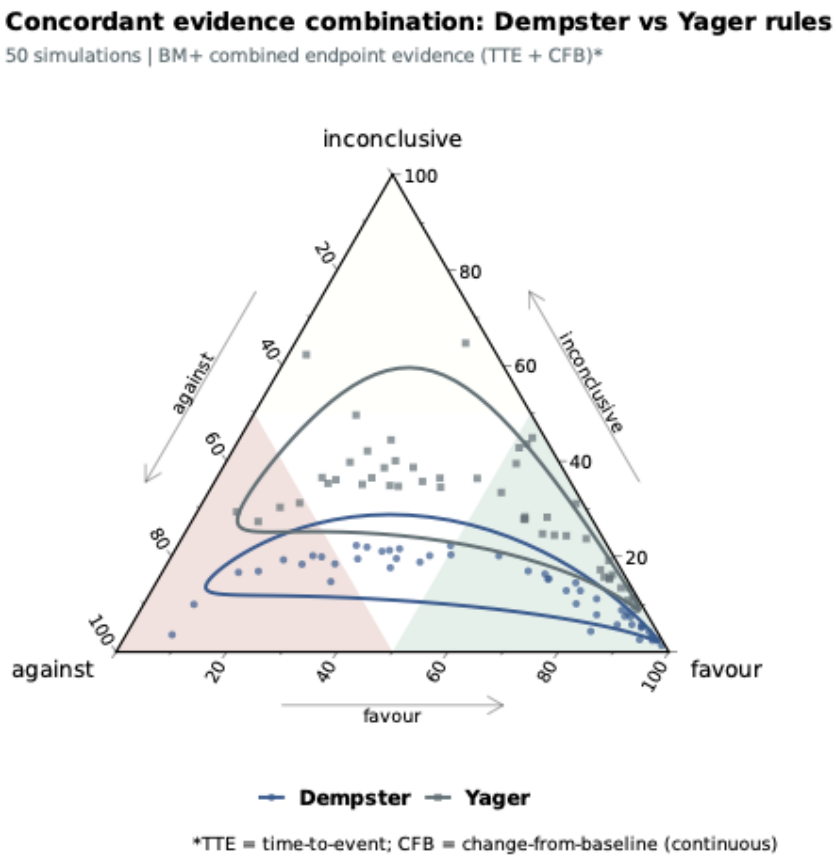
For predictive BM discovery/validation in Phase 2, the false-negative risk far exceeds the false-positive risk

- Detection at HR(BM+)=0.60: Bayesian **40%** vs frequentist **16%**
- Detection at HR(BM+)=0.80 (null): both ≈6–7%
- Same specificity, **2.5× sensitivity**
- → Interaction tests miss most true biomarkers
- *500 sims, N=160, 50% BM+, uninformative prior, Cox PH*



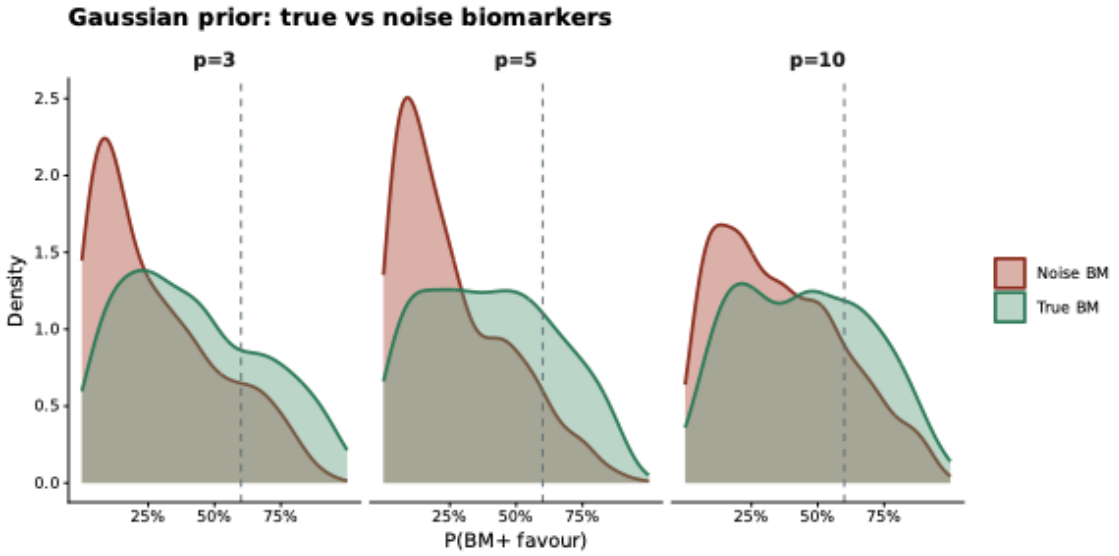
Concordant endpoints amplify the signal via Dempster–Shafer combination

- Both TTE and CFB show BM+ enrichment
- **Individual**: TTE 43%, CFB 47% P(favour)
- **Dempster** (blue): 61% favour, 13% inconclusive
- **Yager** (grey): 50% favour, 31% inconclusive
- *50 sims, N=160; Dempster–Shafer on frame {favour, against, Θ}*



Gaussian priors generalise to biomarker panels as ridge-like shrinkage

- **p** candidate BMs, only 1 truly predictive
- Independent N(0, TPP spread) prior on each BM×Treatment interaction
- At **p=5**: true BM detected **23%**; Bonferroni: **4.5%**
- → **5× more sensitivity** retained
- *Ridge-like shrinkage; noise BMs reflect overall drug effect (~8%)*



A mature framework for precision-medicine decisions in early development

- **2–3× more sensitive** than frequentist interaction testing at comparable specificity — detection of 40% vs 16% at the clinically relevant upside scenario (HR=0.60)
- **Well-calibrated under the null**: false-positive rate of 12.5–15% under the active homogeneous null (HR=0.70 both subgroups), compared to 40% true-positive under the alternative — a signal-to-noise ratio of approximately 3:1 (uninformative prior)
- **Dempster–Shafer evidence combination** amplifies concordant signals across endpoints (43–47% individually → 61% combined) and Yager's rule appropriately retains uncertainty under discordance
- **Ridge-like Gaussian priors** generalise naturally to multi-biomarker panels, retaining 5× more sensitivity than Bonferroni for the truly predictive biomarker

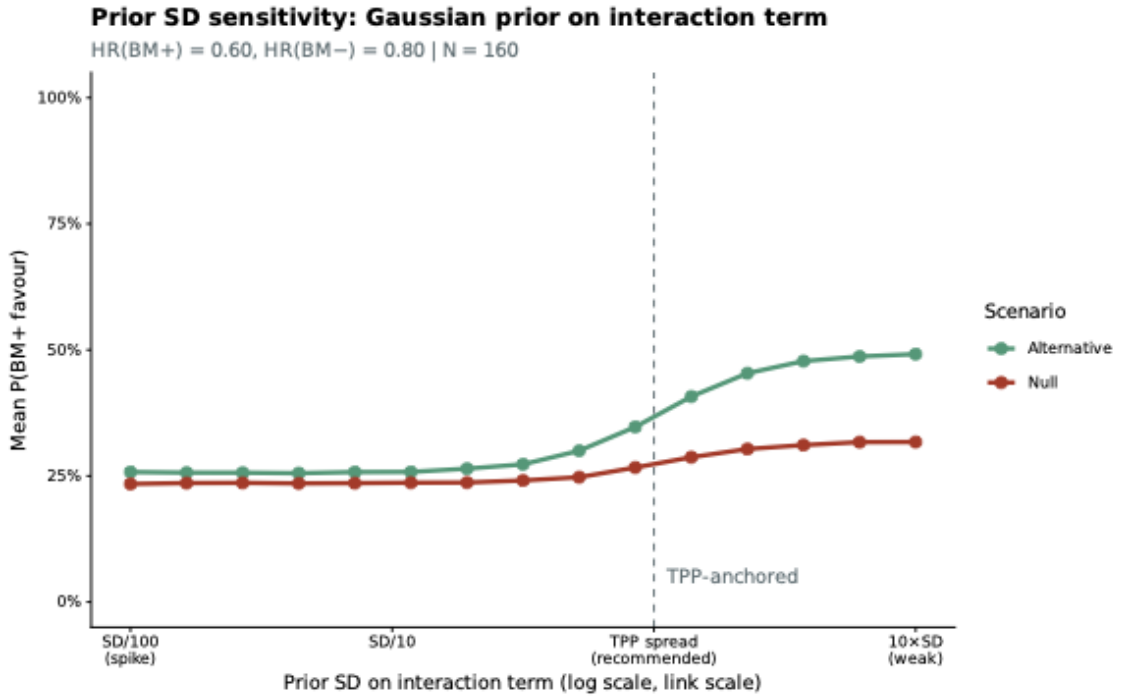
Thank you — questions welcome

Appendix

Backup slides in case of detailed questions

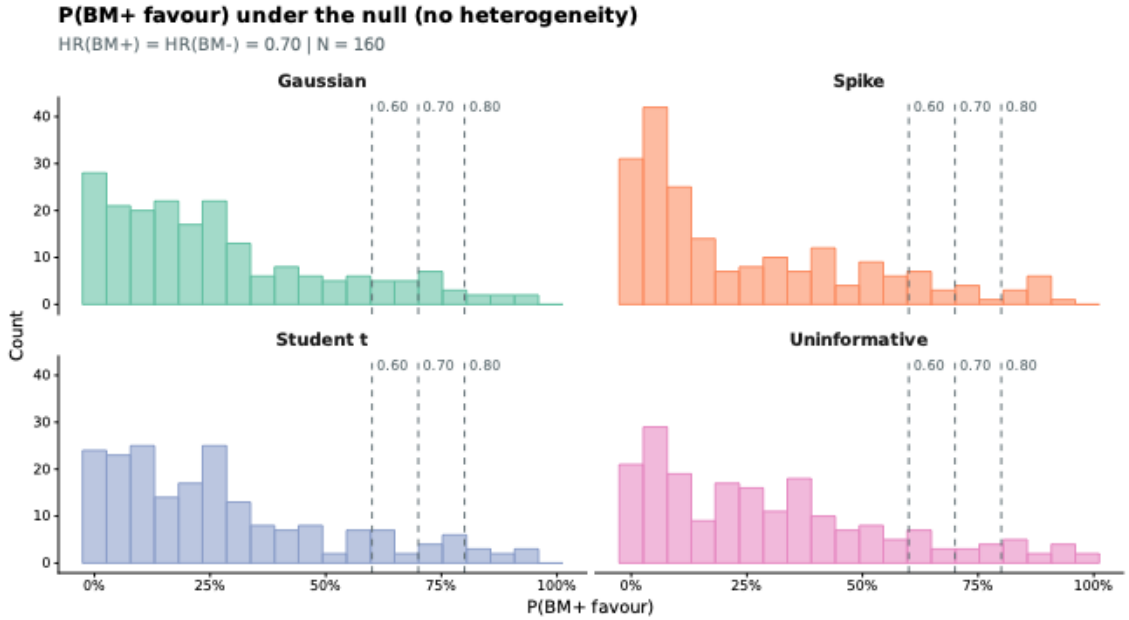
The TPP spread is the natural prior scale — optimal sensitivity/specificity

- Prior SD swept from spike to diffuse (log-HR scale)
- **Green** = detection under alternative; **Red** = FP under null
- Sweet spot at TPP spread ≈ 0.29
- Clinically justified, not arbitrary
- *TPP spread = |log(0.60) – log(0.80)|; 200 sims per point*



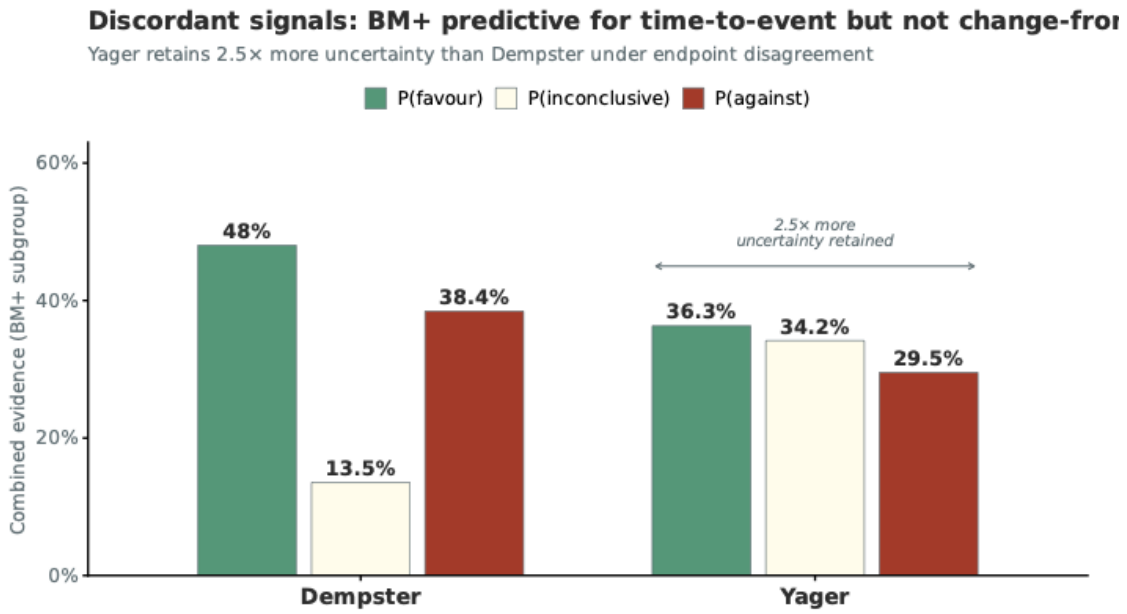
Well-calibrated: 12–15% FP rate under active null at P(favour)≥0.60

- Active null (HR=0.70 both): FP **12.5–15%** at P≥0.60
- Inactive null (HR=1.0): FP **≤1%** (not shown)
- Detection under alternative: **40%**
- Signal-to-noise ratio ≈ **3:1** (uninformative prior)
- *Histogram: P(favour) distribution for BM+ under HR=0.70 null*



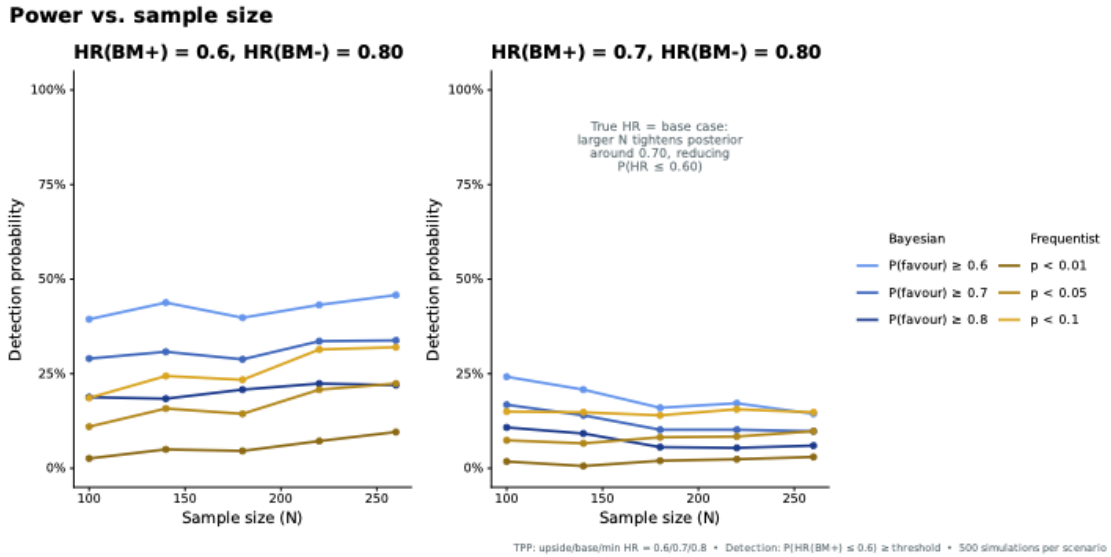
Discordant endpoints: Yager's rule retains uncertainty appropriately

- BM+ predictive for TTE only; endpoints disagree
- Dempster: 48% favour, 13.5% inconclusive
- Yager: 36.3% favour, **34.2% inconclusive** — 2.5× more uncertainty
- Yager is the appropriate default under disagreement
- *Discordant scenario: HR(BM+)=0.60 for TTE; no enrichment for CFB*



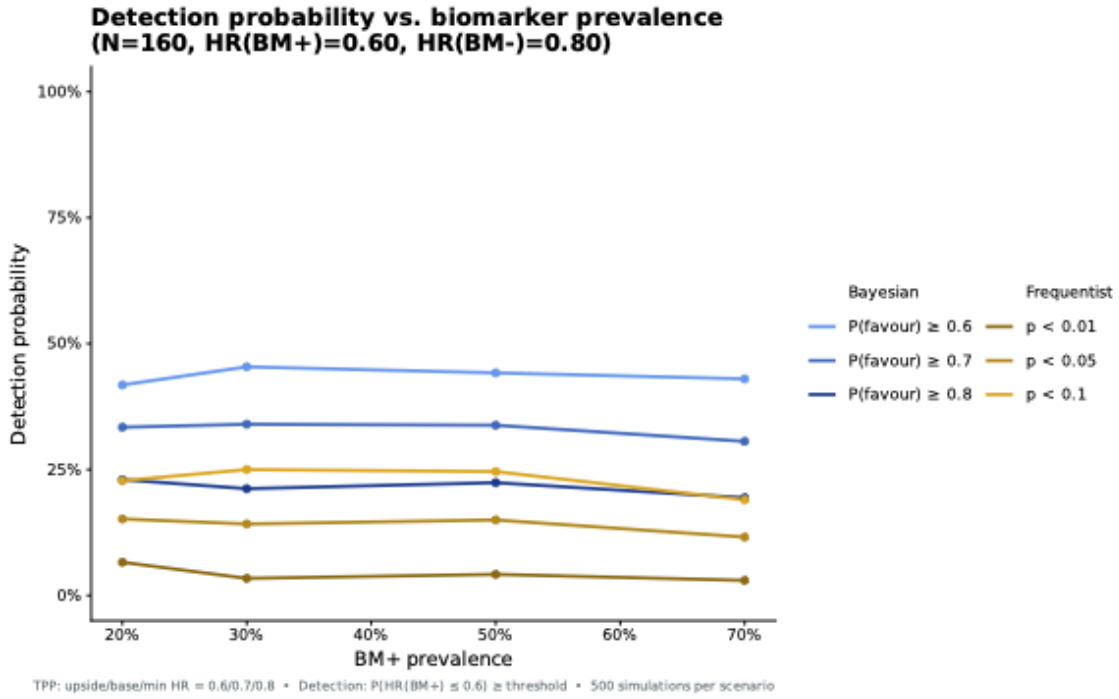
2–3× higher detection than interaction testing, comparable specificity

- **N varied** 100–260 (Phase-2 IQR): strong and moderate signals in BM+ (HR of 0.6 and 0.7)
- Bayesian advantage persists across all sample sizes for HR = 0.6
- BUT when HR = 0.7, increasing sample size means a narrower posterior around the base case, meaning less density in the favour region — so detection actually decreases with N



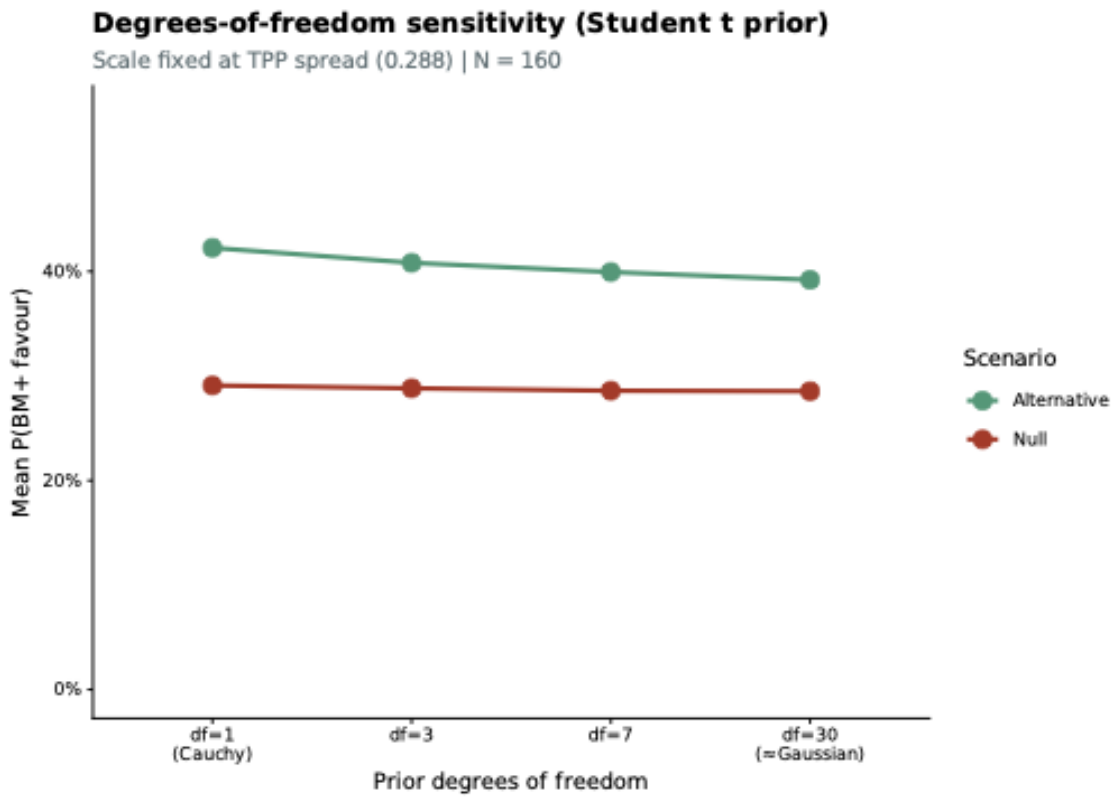
Backup: Detection stable across BM prevalence 20–70%

- **N=160**, HR(BM+)=0.60, HR(BM−)=0.80
- Prevalence varied 20%–70%
- Bayesian P(fav)≥0.6: **39–43%** across all prevalences
- Framework borrows strength through the joint model
- No prevalence-based sample-size adjustment needed



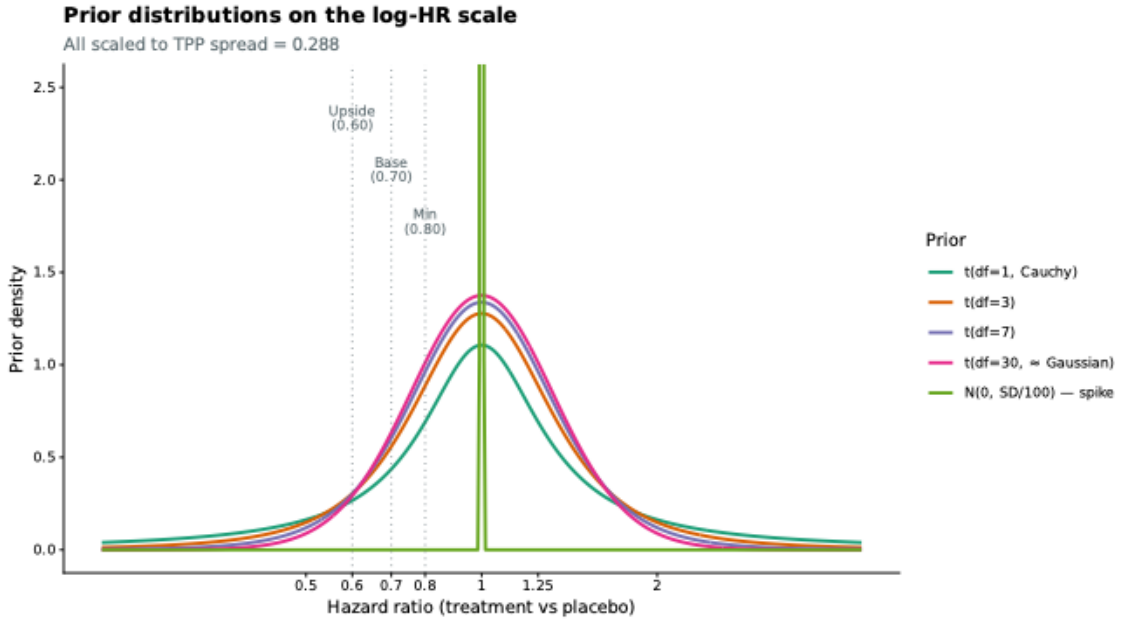
Backup: Conclusions robust across Student-t degrees of freedom

- Scale fixed at TPP spread (0.288)
- df swept: 1 (Cauchy), 3, 7, 30 (≈Gaussian)
- Detection under alternative: **stable** across all df
- Null FP: similarly stable
- → Tail behaviour matters less than scale



Backup: Prior families visualised on the log-HR scale

- All priors centred at 0 (no heterogeneity)
- Scale = TPP spread ≈ 0.288
- Spike: virtually certain no interaction
- Cauchy (df=1): heavy tails permit large interactions
- df=30 ≈ Gaussian: lightest tails
- TPP reference lines show upside, base, minimum



Backup: 5× more sensitivity than Bonferroni for the true biomarker

- Direct comparison: Bayesian vs Bonferroni
- Focus: detection of the TRUE biomarker
- **p=5**: Bayes 23% vs Bonferroni 4.5%
- **p=3**: Bayes 27.5% vs Bonferroni 5.0%
- Bayesian advantage grows with panel size