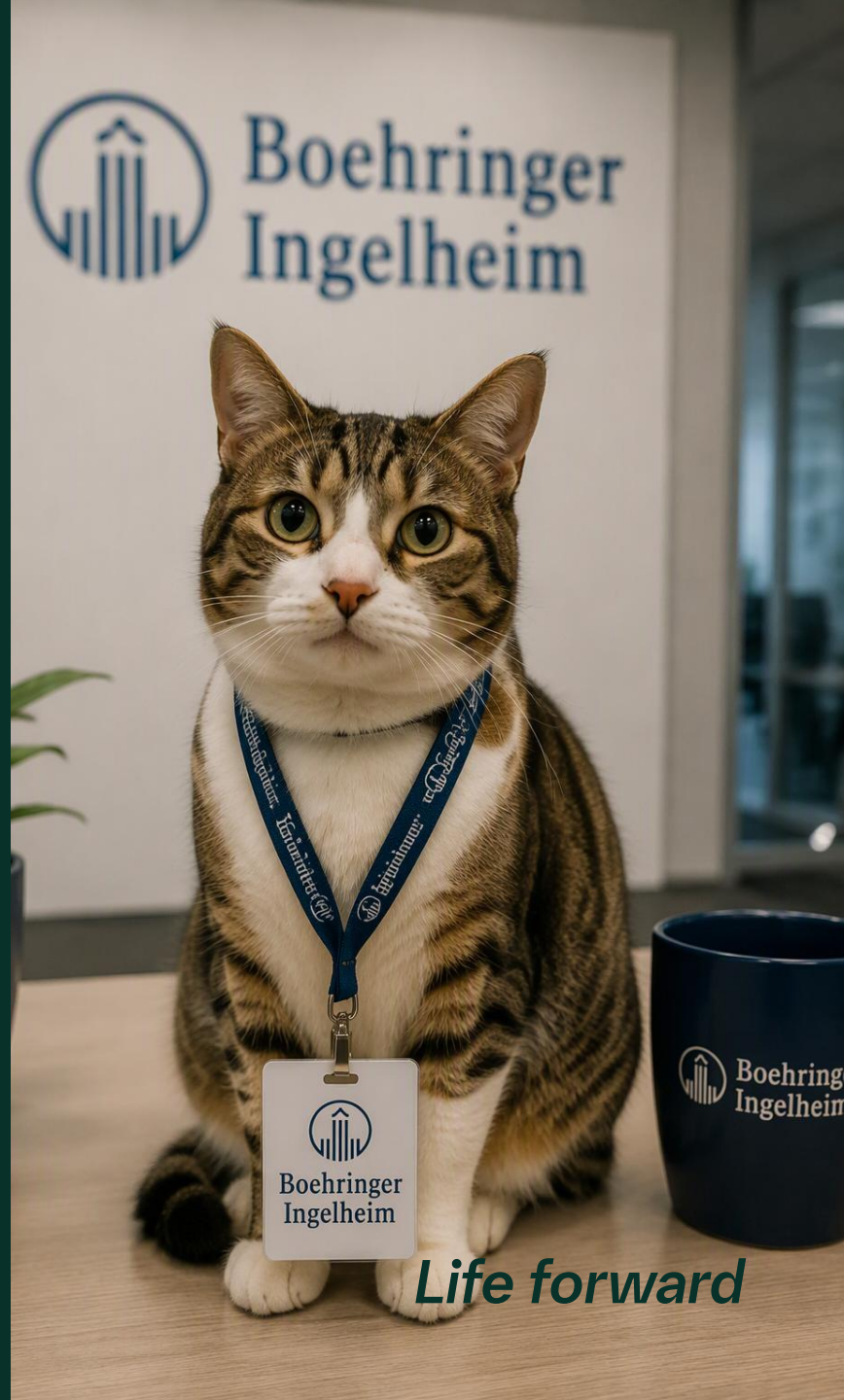
Boehringer Ingelheim

Optional Stopping in Bayesian Logistic Regression

Philip Young & Parsa Mohammadian

Boehringer Ingelheim Vetmedica



Original Study Design

Primary endpoint achieved: upper 97.5% CL 6.3% is below the 15% NI margin

STUDY DESIGN

Open-label non-inferiority trial

1:1 randomized | NI Δ = 15% | FAS · n=116

TREATMENT ARMS

CONTROL
Caninsulin
SC injection · q12h

TREATMENT
Velagliflozin
Oral · q24h

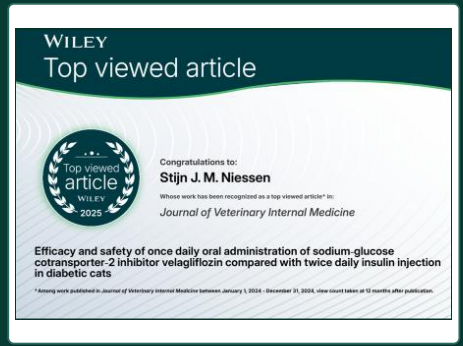
PRIMARY ENDPOINT

Treatment success at Day 45

Binary composite of **Blood glucose** + **Clinical signs**

Secondary endpoints: time points from Day 21

FAS: all randomized cats receiving ≥1 dose with baseline and ≥1 post-Day-21 primary-variable assessment.



TOP VIEWED ARTICLE

Published evidence

Velagliflozin compared with twice-daily insulin injection in diabetic cats

TREATMENT SUCCESS

41.9%	53.7%
Insulin	Velagliflozin
26 / 62	29 / 54

NON-INFERIORITY RESULT

11.8% observed risk difference pT − pC

6.3% upper one-sided 97.5% CL



✓ **NI criterion met: 6.3% < 15%**

Can we do better if we adopt a Bayesian analysis using vague priors?

Bayesian designs are well calibrated to frequentist p-values when:

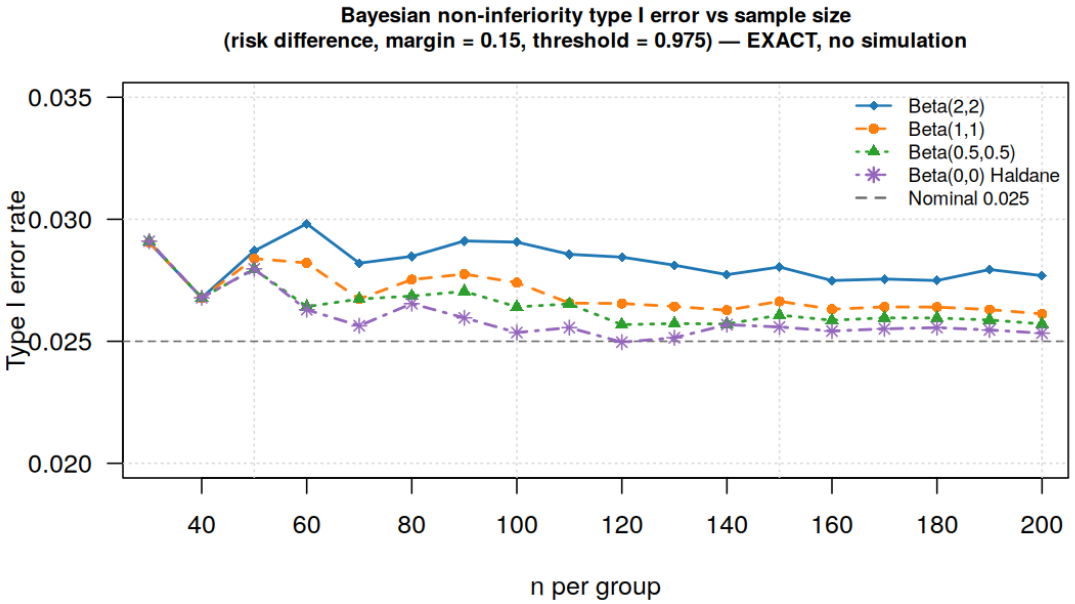
1. **Use a suitable outcome measure**
   The risk difference (RD) is not an optimal outcome measure
   -The odds ratio (OR) is a better measure, translating to a 0.51 OR threshold
   - FDA insisted on using the RD

2. **The Bayes Factor (BF) is used for decision making**
   For the OR the posterior density threshold can be translated to a BF.
   This is not true for the RD because the posterior Odds change
   - FDA prefers the use of the Posterior probability density threshold

3. **A suitable partition of the outcome space is used**
   Point estimates are not good partitions, but non-inferiority is better
   Equivalence does not make sense in this setting
   - Non-inferiority makes sense from a clinical perspective

4. **Classic Bayesian analysis does not need "alpha adjustment"**
   Bayesian "Type I" error averages over the prior and adjusts for interims
   Frequentist "Type I" error uses a worst-case point estimate
   - FDA insists on preserving frequentist-style "Type I" error

Calibration: Using the RD design, small increased type I error



- Power is similar for all four priors
- The sawtooth pattern is real and not an artifact
- Haldane prior fails if 0 or n successes but we can still use it to make deicsions (n>12)
- Study is well powered and Crosses 85% around n ≈ 60 per group
- Power will decrease if we adjust for α-inflation
  - Haldane and Jeffrey have almost no adjustment
  - Agresti leads to loss ~ 6% power

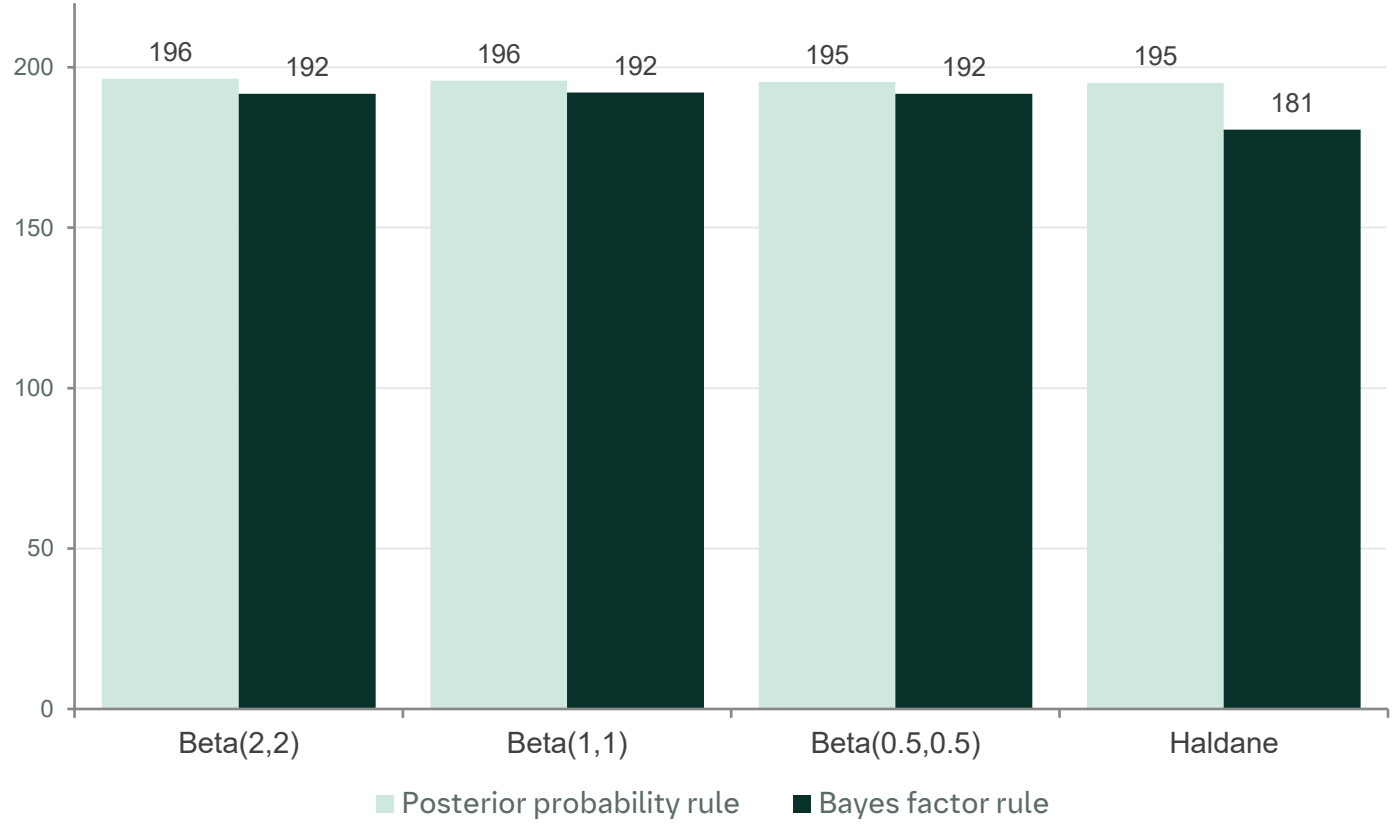
Boehringer Ingelheim

Stopping Rules for Bayesian Non-Inferiority

Two decision rules — posterior probability threshold vs. Bayes factor — calibrated per prior to exactly α = 2.5% type I error

Group sequential design · looks at n = 50, 100, 150, 200 per group · non-inferiority margin Δ = 0.15

Expected sample size under the null (per group)



2.5% Type I error — identical for both rules, all four priors, by construction of the calibration

Prior	Overall Power	Calibrated BF
Beta(0,0) Haldane	99.9%	**37.8**
Beta(2,2)	99.9%	**59.0**
Beta(1,1)	99.9%	**66.9**
Beta(0.5,0.5)	99.9%	**68.1**

What calibration reveals

- Posterior-probability rule needs 0.975 → ~0.991–0.992; Bayes factor thresholds range 38–68 depending on prior — both corrections are prior-specific, not one-size-fits-all.
- Haldane + Bayes factor is the standout: stops ~15 patients per group earlier than any other combination when the null is true, at no cost to power.
- Overall power is identical at fourth interim, but most of the action occurs at first and second interims – so we designed a smaller study
- **Although BF is optimal, we use posterior probability threshold and use Jeffreys prior as it has higher stability**

Boehringer Ingelheim

Optimal stopping rule: two-arm risk difference

H0: pT ≤ pC · interims at 24 and 46 per arm, final at 60 · power ≥ 85% · interim cost = 10 patients · min 10 patients between interims

Stopping boundaries (control ≈ 30% responders)

n per arm	Control responders	Futility: treatment ≤	Efficacy: treatment ≥
24	7	9	16
46	14	20	26
60 (final)	18	29	30

Operating characteristics

Metric	Optimal sequential	Fixed, 52 per arm
Max type I error	2.1% (2.4% non-binding)	2.5%
Power, 30% vs 60%	85.5%	85.7%
Expected total N, null	58.6	104
Expected total N, alternative	78.4	104
Expected interims, null / alternative	1.2 / 1.6	0
Maximum total N	120	104

-44%
Expected N if no difference

-25%
Expected N if RD = 0.30

79%
Null trials stopped at first interim

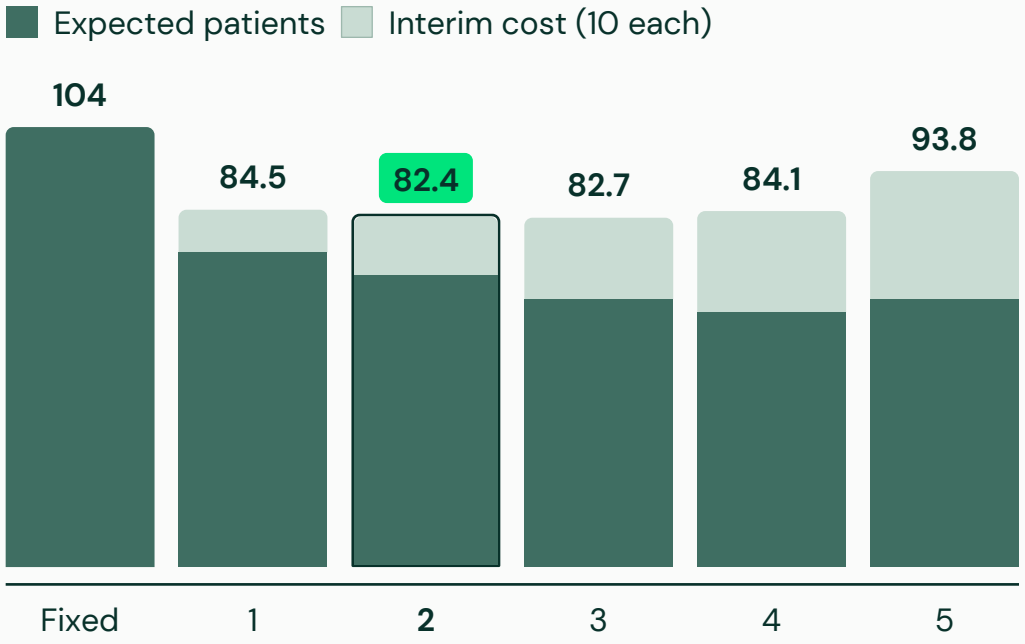
Boehringer Ingelheim

Exact calculations; Jeffreys Prior: Cut-offs rise by about 1 per extra control responder. Type I error is the maximum over common response rates 0.05–0.9; it rises to 2.4% if both arms are near 95%.

How many interim analyses?

With each interim costing 10 patients: interims at 24 and 46 per arm, final at 60

Total cost by number of interims



Number of interims, best timing for each · highlighted = recommended (lowest total cost)

Best timing for each option

Interims	Looks per arm	Exp. N	Exp. interims	Total cost
Fixed	52	104	0	104
1	29, 54	74.5	1.0	84.5
2	24, 46, 60	68.5	1.4	82.4
3	16, 32, 44, 60	63.5	1.9	82.7
4	12, 24, 34, 50, 60	60.1	2.4	84.1
5	10, 20, …, 60	63.3	3.0	93.8

1 to 4 interims are within about 2 patients of each other, so other factors can decide. One interim keeps the maximum at 108 instead of 120.

Boehringer Ingelheim

Total cost = expected total N + 10 × expected interims performed, averaged over null and alternative · looks at least 10 per arm apart, max 60 per arm · type I ≤ 2.5% (futility non-binding) · power ≥ 85%

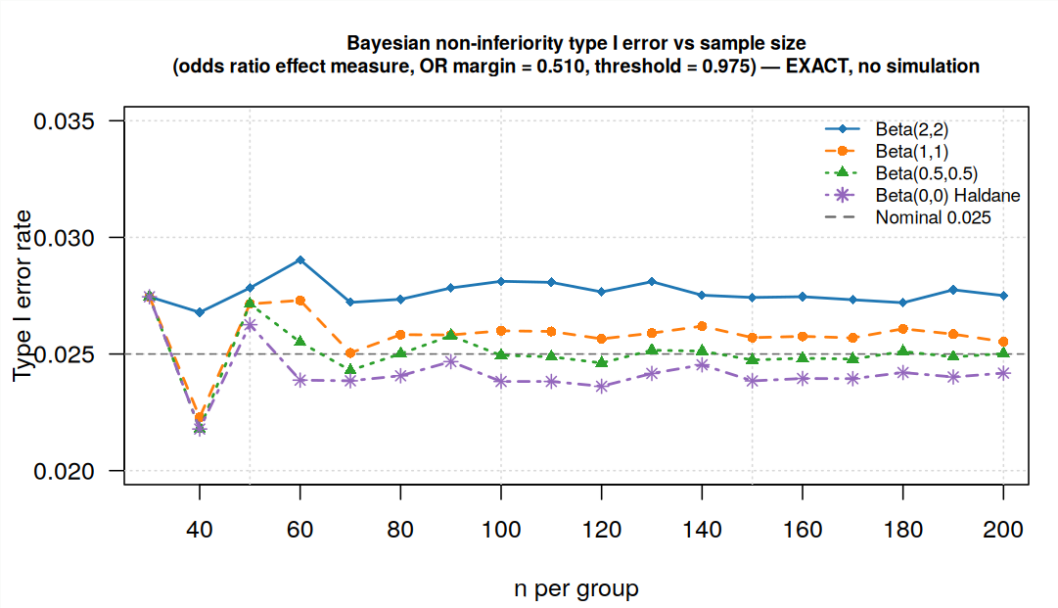
Risk difference vs odds ratio

Same hypothesis (pT > pC is OR > 1), Different Priors; Posterior Density vs BF; After calibration operating characteristics very similar

Recommended designs side by side

	RD design	OR design
Prior	Jeffreys	Haldane
Final-look threshold, P(pT > pC)	0.982	0.984
Looks per arm	24, 46, 60	20, 37, 60
Expected N, null / alternative	58.6 / 78.4	54.6 / 80.1
Expected interims	1.4	1.5
Total cost	82.4	82.5
Power, 30% vs 60%	85.5%	85.0%
Max type I error (non-binding)	2.4%	2.4%

OR design: calibration of type I error not needed



Both designs: interim cost 10 patients, looks at least 10 per arm apart, max 60 per arm, type I ≤ 2.5% for any common response rate (futility non-binding), power ≥ 85% at 30% vs 60%

Boehringer Ingelheim

Questions?

Boehringer Ingelheim