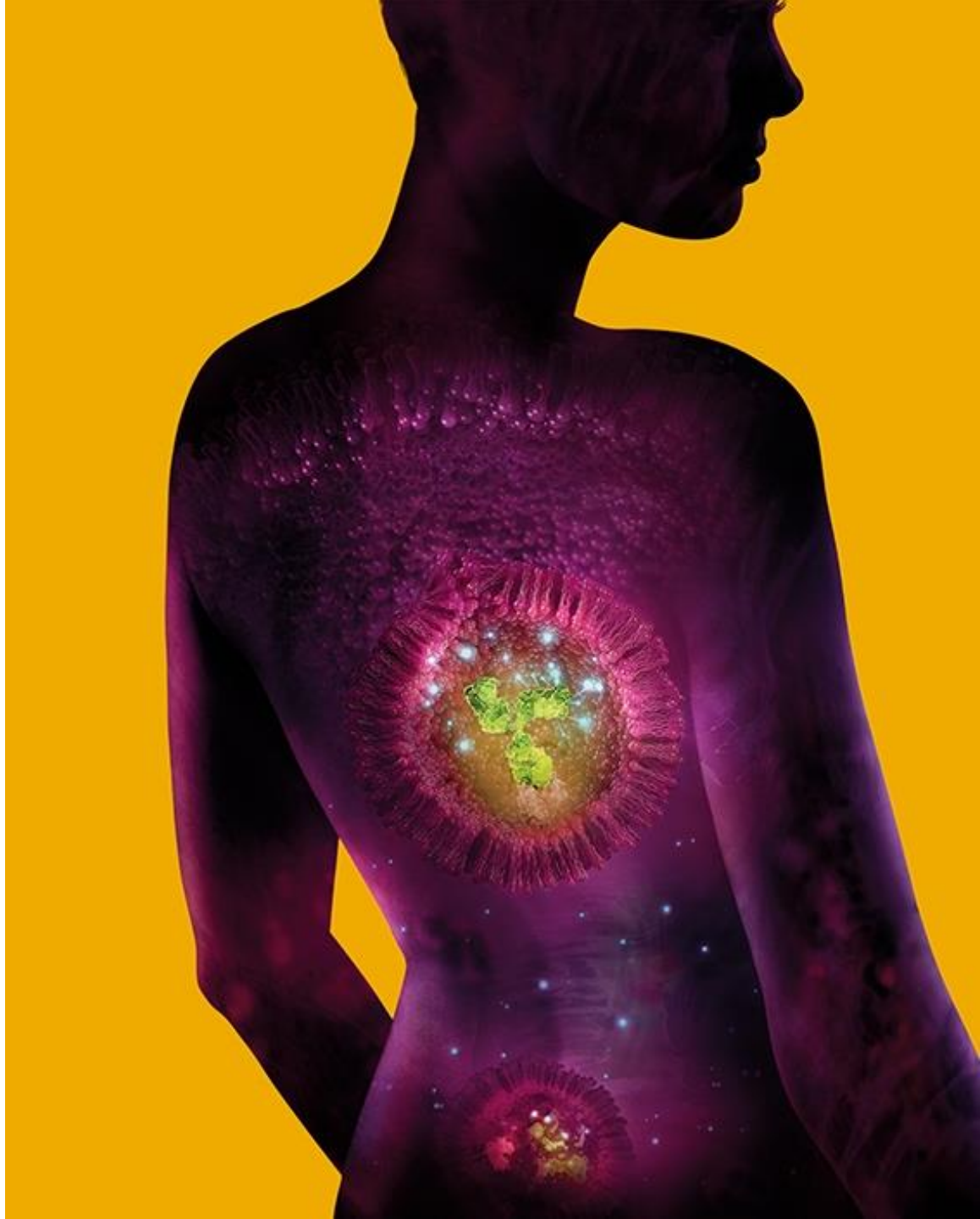
What's hot in pharma statistics?

Carl-Fredrik "Caffe" Burman,

Statistical Innovation, Biopharmaceutical R&D

AstraZeneca R&D Gothenburg

ISBS, November 7, 2025



Background: Requirements to get marketing authorisation: A (not entirely true) summary

- Demonstrate that drug X has positive efficacy
  - Compare vs Placebo
  - Often X vs Placebo on top of standard-of-care placebo+standardbehandling
  - Sometimes indirect comparison (when placebo is unethical)
- Pre-specify “everything”
- Statistical significant efficacy, α=2.5% (one-sided), in two trials
- (US:) Everything in label (cf. FASS) must be “proved” at multiple test family-wise error rate α=2.5%
- Safety, benefit/risk, quality …

A few milestones ..

- FDA first requiring proof of efficacy (1962)
- Group-sequential designs (1970s)
- Multiple inference (Holm, 1979; Hochberg, 1988; ICH E9, 1998)
- Adaptive designs (Bauer & Köhne, 1994; trending a decade later)
- Human genome project (2000; personalised medicine (F Collins), precision med)
- …

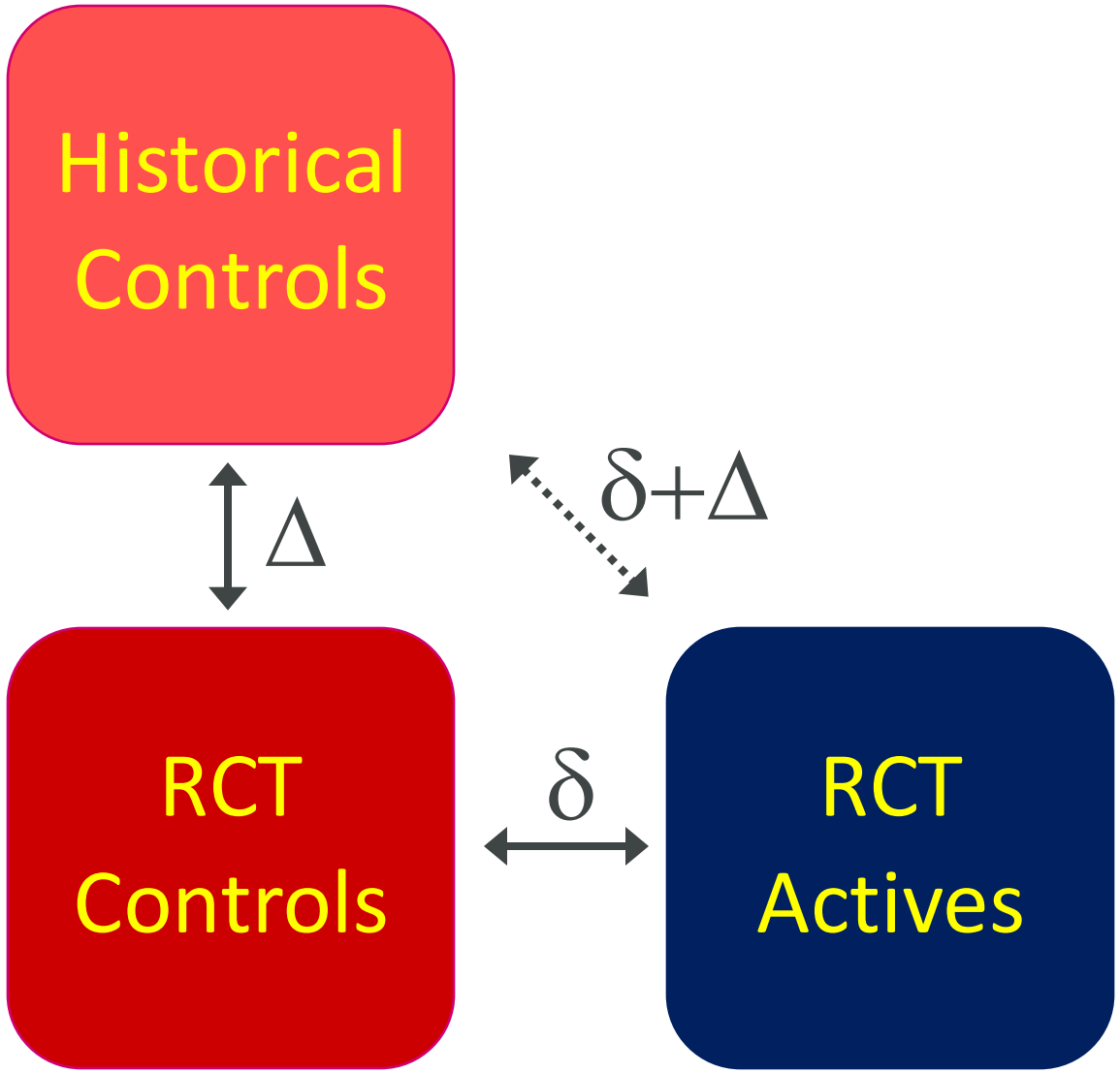
… and areas to mention

- Rare diseases
- Pharmacometrics; Supporting evidence
- Estimands
- RWE
- Master protocols
- Bayesian borrowing
- AI/ML
- PROCOVA
- What else?

Master protocols

- FDA guidance (draft 2023)
- Basket
- Platform
- Core ideas: perpetual infrastructure; shared control; modular add/drop arms
- Impact: reduced setup time, comparable evidence across related therapies
- Hot topics: control arm drift, non-concurrency adjustments, multiplicity

Bayesian (Dynamic) Borrowing



Estimands

- ICH E9(R1)
- Define: treatment, population, variable, intercurrent event strategy, summary measure
- What's new: strategy choices (treatment policy, hypothetical, composite, while-on-treatment)
- Hot topics: aligning design, data collection, and analysis with the chosen estimand

AI/ML

- Where it helps: signal detection, risk-based monitoring, image/omics endpoints
- Analysis vs. product: exploratory decision support vs. prespecified confirmatory analyses
- Guardrails: model governance, fairness/robustness

Bayesian decision-making

- Concept: utility/risk frameworks across program, not just trials
- Metrics: probability of technical/regulatory/Commercial success, VOI, Go/No-Go
- Practice: prior elicitation, cross-study synthesis, portfolio simulation
- Lalonde

Design
is an investment decision

Design
lives in a context

Design
parsimony is a virtue

Design
good > ”innovative”

Design
trial → programme

Design
no trial may be best

EXAMPLE. Biomarker-defined sub-population

Pop.	True Effect	SD	Inclusion Rate (pat./month)	Trial Cost (MUSD)	Interim Cost (MUSD)	Commercial Value (MUSD)
Full pop	0.205	1.0	100	10 + 0.02*N	3 + 1*a	(2000-25*T)*TE
BM+	0.273	1.0	40	10 + 0.025*N	3 + 1*a	(1000-10*T)*TE
BM-	0.160	1.0	60			

Design option	Adaptation	N	Inclusion time	Trial Cost	Power			PoS	True eNPV (MUSD)
					Any	AC	BM+	Any	
Full pop	None	1000	10.0	30	90%	90%		77%	246
BM+	None	563	14.1	24	90%		90%	77%	157
AD	Stage 2 pop	1000	10.0 (17.5)	33	94%	79%	15%	80%	237

TE=True effect, BM+/-=Biomarker positive/negative
AD=Adaptive design with subgroup selection at interim and the same total sample size
This example is just BASED on a true story. Also, assurance should be included.

Multiple testing in A and B

- Common to test null hypothesis of no efficacy in A, and null hyp of no efficacy in B
- Control Family-Wise Error Rate at $\alpha = 2.5\%$:
  - May test B first, test A iff B is stat sign
  - Or the other way around: A followed by B
  - Or split alpha between A and B (Spiessen-Debois utilizes correlation), followed by recycling if any hypothesis is stat sign.

- What does it mean?
  - We first conclude that Active drug is better than Placebo in subpopulation B, say
  - Then we test if there is any efficacy on average vs. Placebo in $A = B \cup C$
  - How could that not be true?

The logic of rejections

- Winning in A does not imply any efficacy in C,
- but testing separately in C often lacks power
- Borrowing efficacy data between B and C? We'll look at that later.

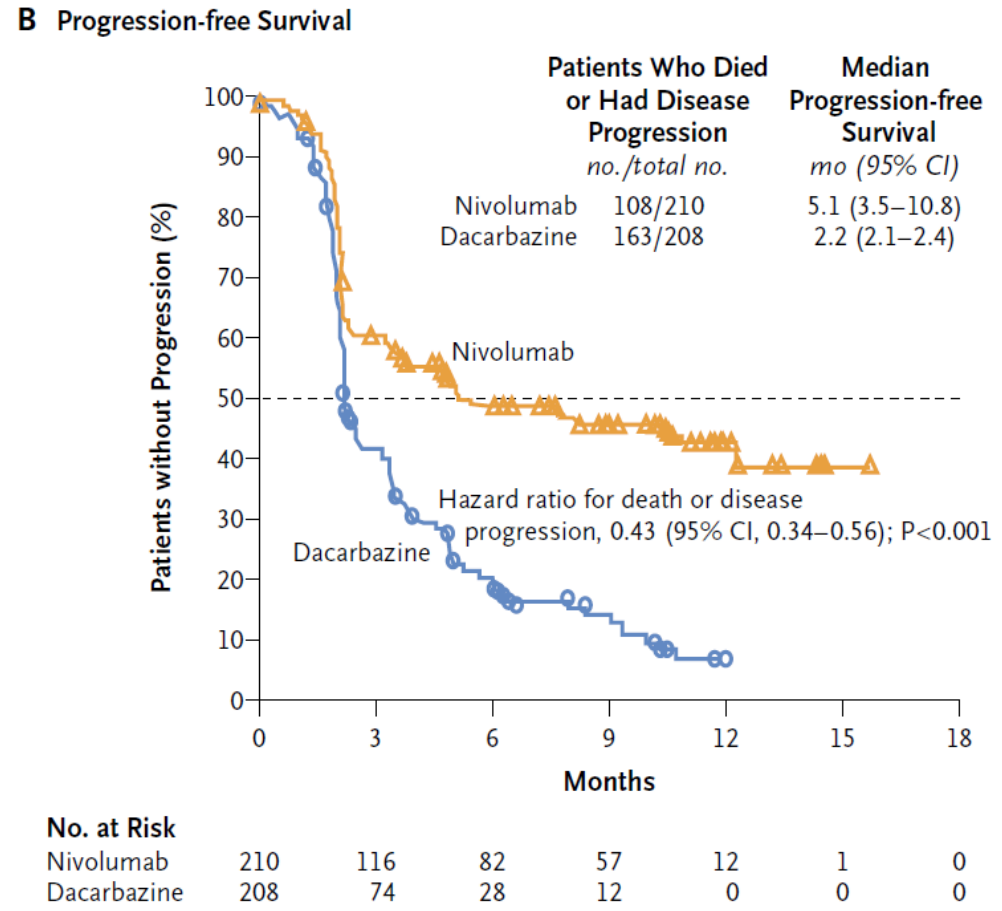
- In general, a statistical significance (if not a chance finding) in A, say, only implies
  - There exist patient(s) in the RCT population who have positive efficacy
- Extrapolation from RCT to the real world relies on extra-statistical assumptions

What should regulators require?

1. “Proof” of efficacy vs. Placebo
   - Relevant endpoint (how the patient feels, functions or survives)
   - Preferably P-value < 0.1% (cf. 2-trial rule with α=2.5%)
   - Relax this when enough power isn’t feasible
   - Bayesian: May be strengthened by supportive evidence, Borrowing
   - Focus on hypothesis testing, validity & power, not on estimands and estimates
   - Cannot conclude that everyone benefits
2. Likely positive benefit/risk in a certain (sub)population in clinical practise
   - Cannot normally “prove” benefits > safety risks
   - Focus on estimates, right estimands
   - Need modelling, assumptions, extrapolations, expert assessment

Extra: Nonproportional hazards (NPH)

- Many trials of PD1 / PDL1 drugs vs chemotherapy showing better efficacy over time



Weighted logrank test

$$T = \sum_{j=1}^{k} w_j \left(d_{0,j} - d_j \frac{n_{0,j}}{n_j} \right)$$

- Fleming-Harrington(ρ,γ): $w(t) = S(t)^{\rho} \cdot F(t)^{\gamma}$
- FH$(0,1)$ literally gives weight zero to the first death
- MaxCombo: Omnibus test with different weighting schemes
- MaxCombo and F(0,1) controls the T1E in a “two-sided” test but not in the correct “one-sided” test

Survival function S(t) and CDF F(t) above are estimated from pooled data

What do we “prove” with a hypothesis test?

$$H_0^{\text{weak}}: \quad S_1(\text{t}) = S_0(t) \quad \text{for all } t$$

$$H_0^{\text{strong}}: \quad S_1(\text{t}) \leq S_0(t) \quad \text{for all } t$$

Magirr & Burman: “The MaxCombo test severely violates the Type 1 Error”, Letter in JAMA Oncology (2023)

Bayesian Dynamic Borrowing of control group data

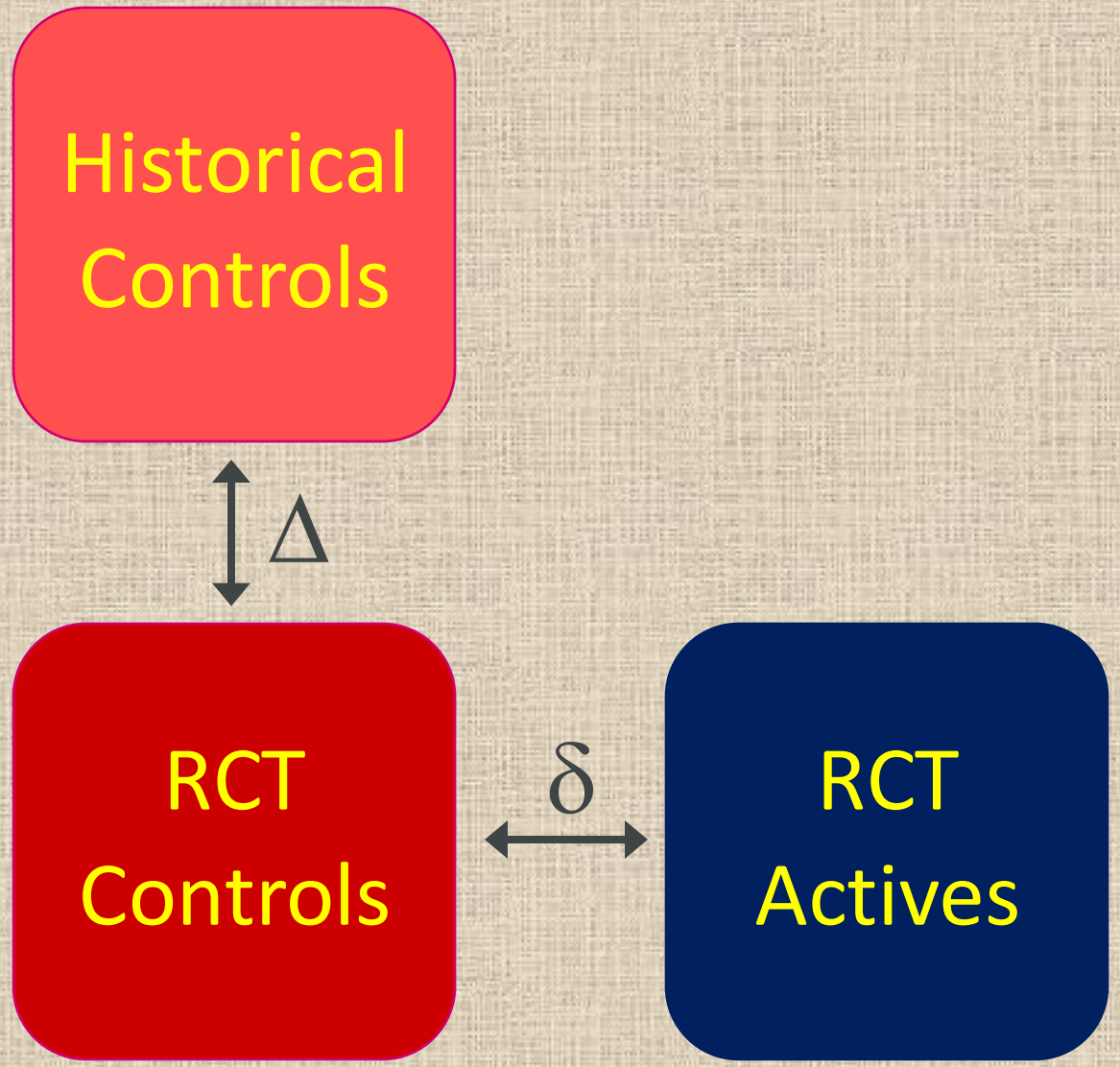
Pharmaceutical Statistics
The Journal of Applied Statistics in the Pharmaceutical Industry
The official journal of PSI

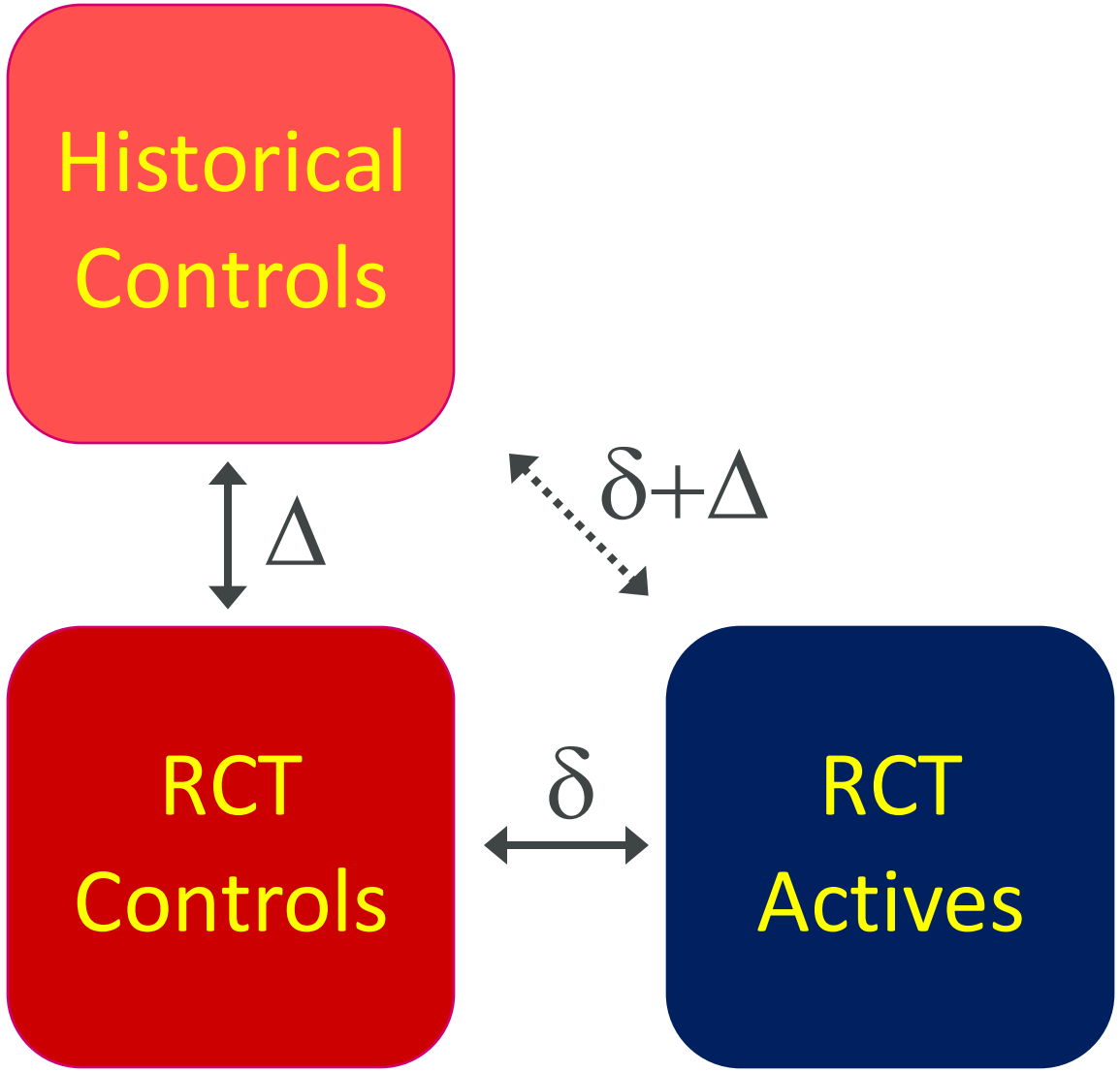
MAIN PAPER | Full Access

Digital twins and Bayesian dynamic borrowing: Two recent approaches for incorporating historical control data

Carl-Fredrik Burman, Erik Hermansson, David Bock, Stefan Franzén, David Svensson

First published: 04 March 2024 | https://doi.org/10.1002/pst.2376




Historical Controls
Δ
δ+Δ
RCT Controls
δ
RCT Actives

Bayesian Dynamic Borrowing (BDB): A simple variant

- Optimistic scenario
  - No mean difference TC vs HC, $\Delta=0$
  - Prior is certain that $\Delta=0$. Data cannot change this, Posterior also has $\Delta=0$ with probability 1.
  - Historical controls (HC) and trial controls (TC) are pooled.
- Pessimistic scenario
  - HC fairly unrelated to TC, large uncertainty in Δ.
  - Information in data overwhelms prior. Posterior for Δ is approximately $N(\widehat{\Delta}, 2\,S^2/N)$.
  - HC are essentially disregarded.
- Compromise (dynamic prior)
  - We'll see later how the weight is updated

Example of posterior weight

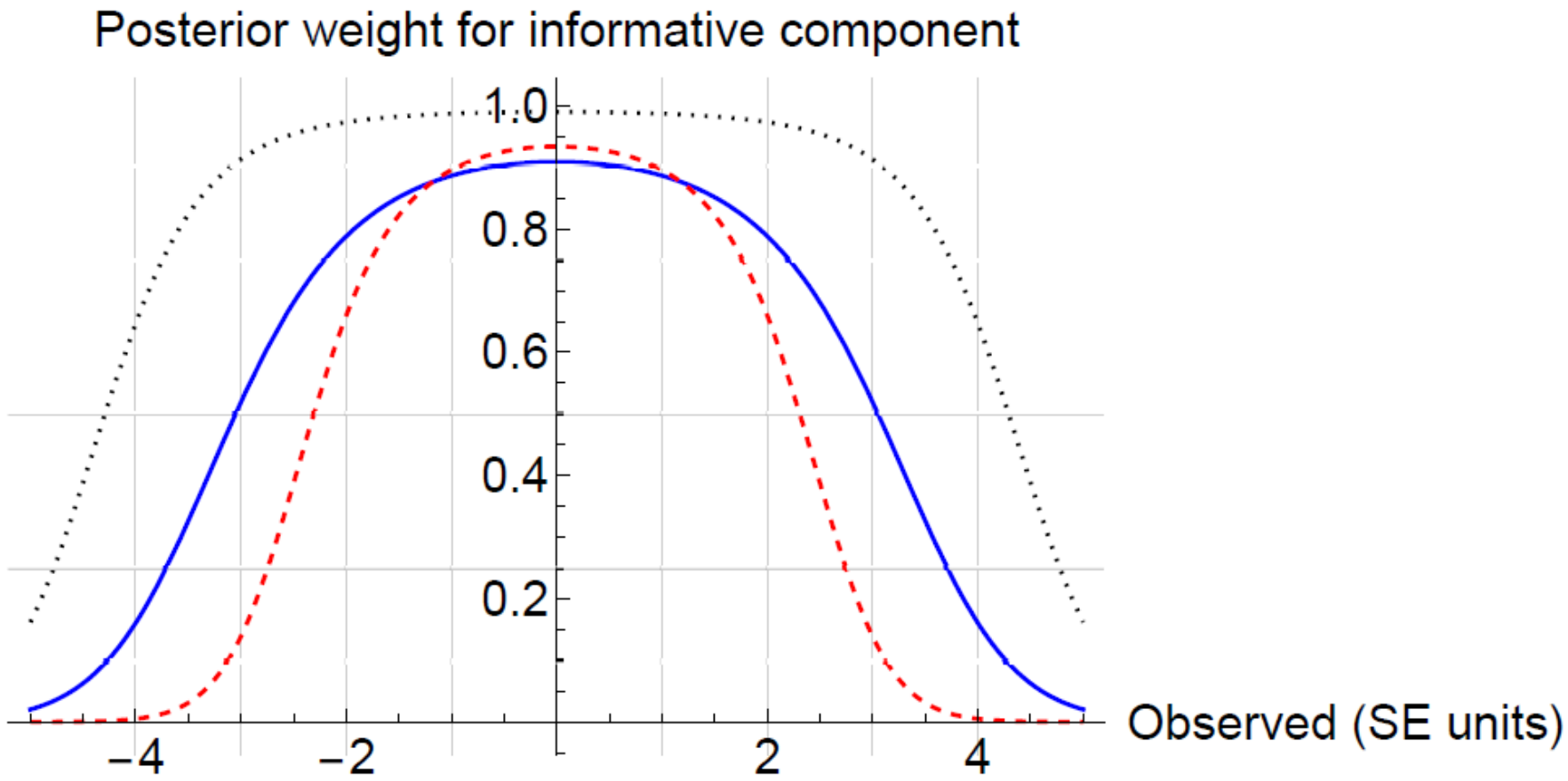


Figure: Weight for the informative component, as function of $\hat{\Delta}/SE[\hat{\Delta}]$. Solid blue curve has $n^\star = n = 200$. The dotted black curve results when the vague prior has ten times larger standard deviation. Finally, the red dashed curve use the same model as the blue but with $n^* = \infty$.

Posterior weight is pretty random!

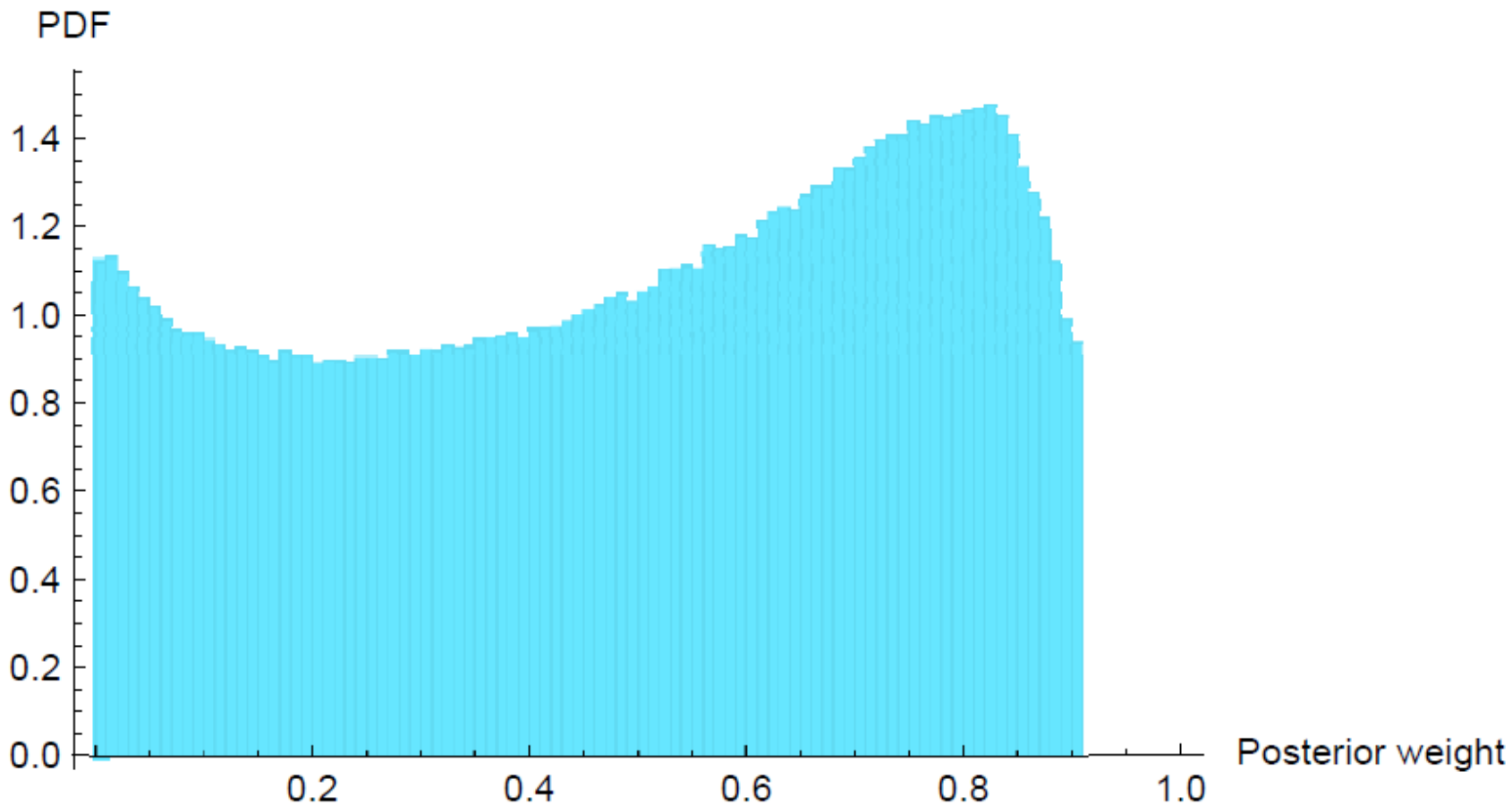
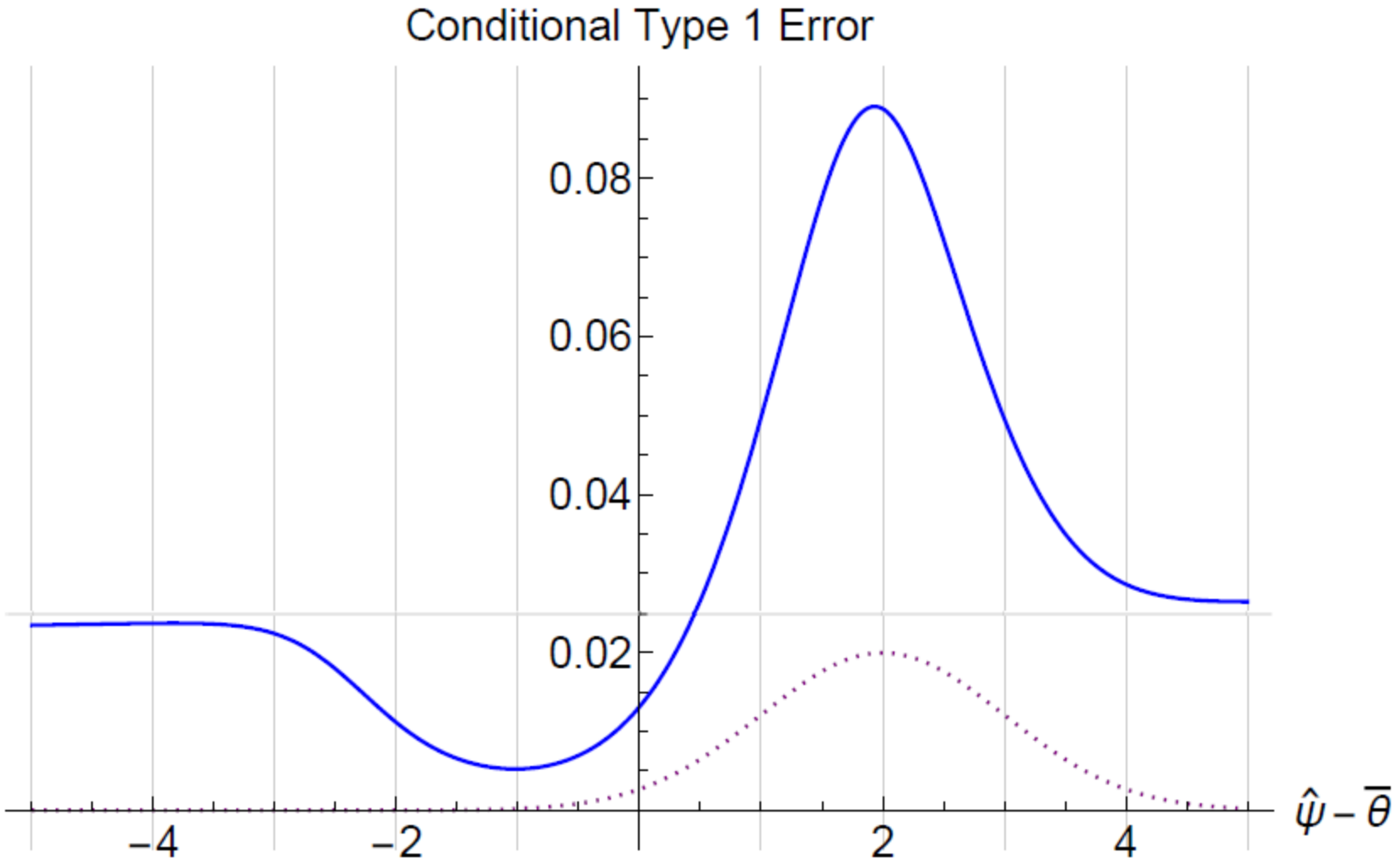


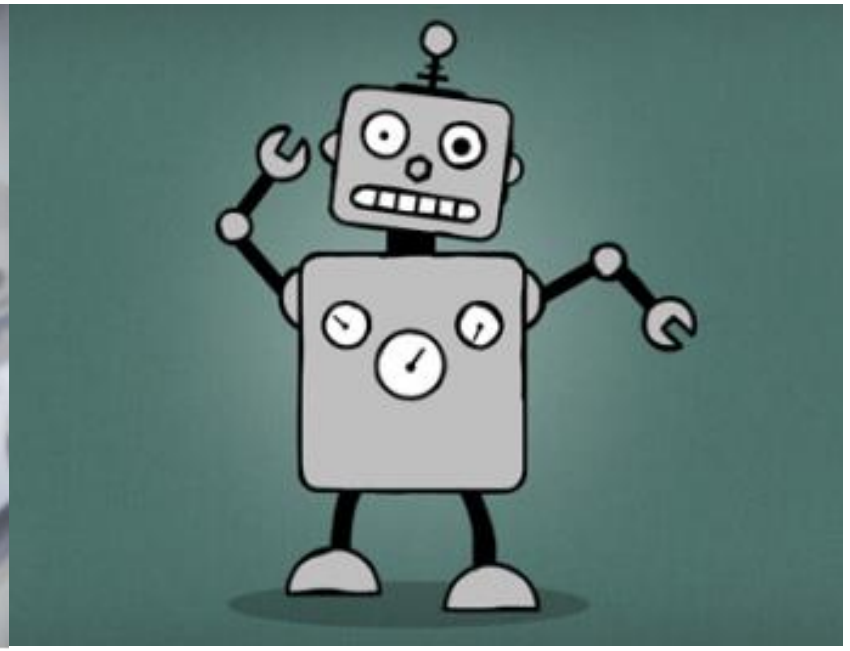
Figure: Probability density for posterior weight when RCT control mean is 3 standard error units away from HC average.

Very large Type 1 Error inflation



Prognostic Score Methodology

What if the doctor could guess the outcome?



What if an AI could guess the outcome?

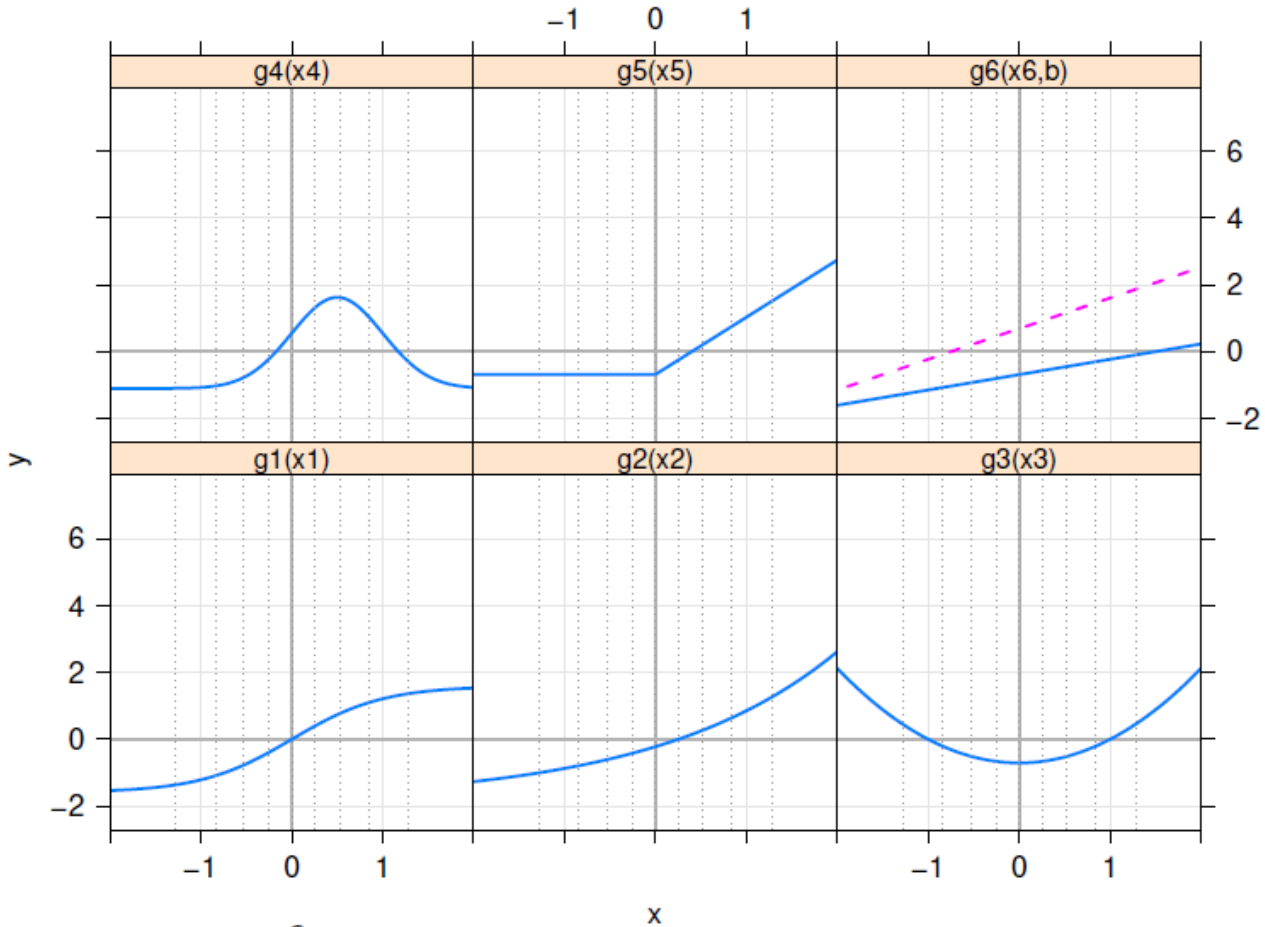
Train Random Forest on historical data set $Y_i(x_i)$

Use the RF predictor z(x) for each patient as covariate in RCT

DT simulation set-up

$$y(\mathbf{x}, \mu) = \mu + g_1(x_1) + g_2(x_2) + g_3(x_3) + g_4(x_4) + g_5(x_5) + g_6(x_6, b) + \varepsilon$$

where $x_j \sim N(0, 1)$ independent



$$\mathbf{Var}(y) = \sum_{j=1}^{6} \mathbf{Var}(g_j(\mathbf{x})) + \mathbf{Var}(\varepsilon) = 6 \cdot 1 + 4 = 10.$$

DT simulation results

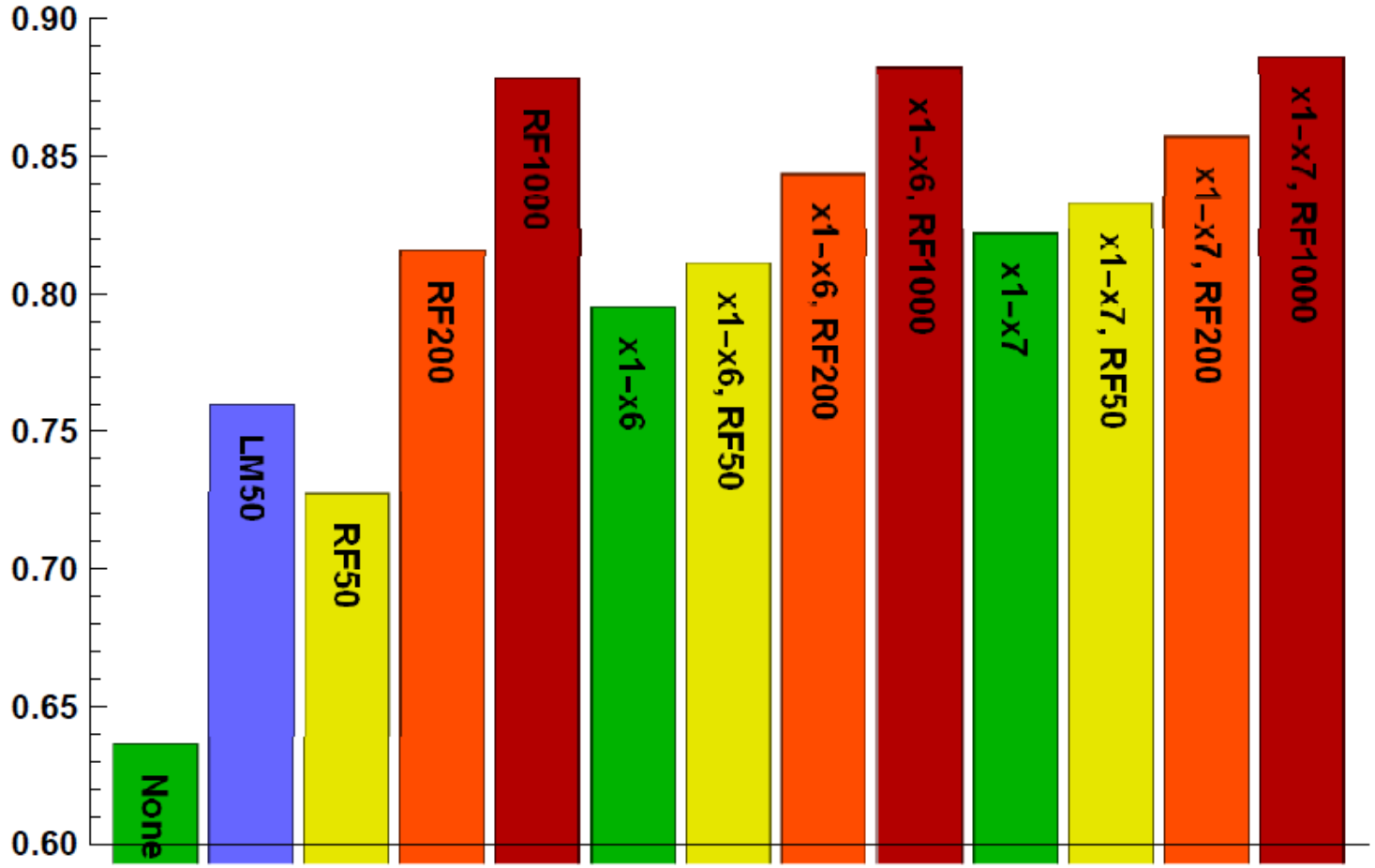


Figure: Power for ANCOVA models with 0, 6 or 7 baseline covariates (green); similar models complemented by an RF-trained predictor, based on $n^* = 50$ (yellow), $n^* = 200$ (orange) or $n^* = 1000$ (dark red); or complemented by a linear predictor ($n^* = 50$, blue) based on historical data.

Discussion

Bofill Roig *et al. BMC Medical Research Methodology* (2022) 22:228
https://doi.org/10.1186/s12874-022-01683-w

BMC Medical Research Methodology

RESEARCH Open Access

On model-based time trend adjustments in platform trials with non-concurrent controls

Marta Bofill Roig[1], Pavla Krotka[1], Carl-Fredrik Burman[2], Ekkehard Glimm[3,4], Stefan M. Gold[5,6,7], Katharina Hees[8], Peter Jacko[9,10], Franz Koenig[1], Dominic Magirr[3], Peter Mesenbrink[11], Kert Viele[12] and Martin Posch[1*]

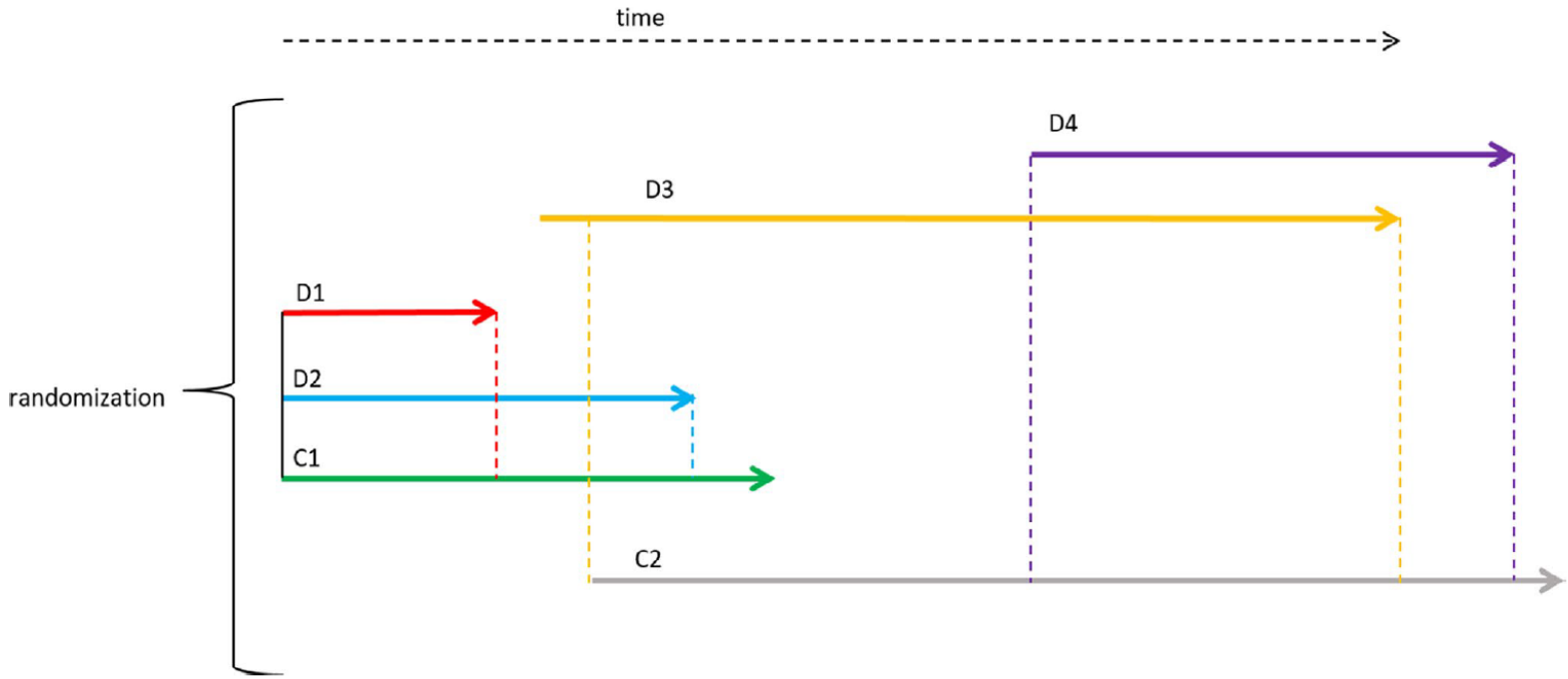


Figure 5 The platform trial starts as a three-arm randomized trial including drugs D_1, D_2, and an active comparator C_1. As data accrue the treatment arms D_3 and D_4 and another active comparator C_2 are added. [Colour figure can be viewed at wileyonlinelibrary.com]

Non-concurrent controls (NCC)

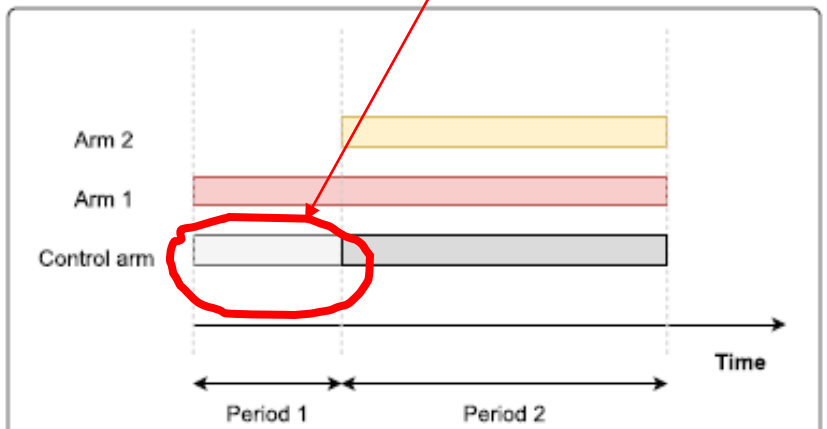


Fig. 1 Platform trial scheme. Example platform trial with initially a treatment group (arm 1) and a control group in period 1; and with a new treatment (arm 2) starting in period 2. Light grey represents non-concurrent controls with respect to the new treatment, dark grey represents concurrent controls

Remarks

- There are limits to what RCTs can do
- Use all relevant data (~sufficiency principle)
- Borrowing makes sense when needed
- It’s all about good statistical principles!
- Include science
- Explicit assumptions
- Ask the right precise question(s) …
- … and tailor the answers
- Check robustness
- The greatest flaw with modern Bayesian statistics is that it isn’t enough Bayesian

AstraZeneca PLC, 1 Francis Crick Avenue, Cambridge Biomedical Campus, Cambridge, CB2 0AA, UK
+44(0)203 749 5000
www.astrazeneca.com