Joint DSBS/FMS meeting
5 November 2024
Kastrup Strandpark

DSBS Danish Society for Biopharmaceutical Statistics

FMS Swedish Society for Medical Statistics

This year’s Organizing Committee

- Carl-Fredrik Burman, AstraZeneca
- Jonathan Bergman, Läkemedelsverket
- Ketil Biering Tvermosegaard, Novo Nordisk
- Kyle Raymond, Ferring Pharmaceuticals
- Mads Wessel Pedersen, Lundbeck
- Mikala Fiig Jarner, LEO Pharma
- Thor Schütt Svane Nielsen, Zealand Pharma

Today's programme

Timeslot	Speaker	Title
8:30 – 9:00		Breakfast and arrival
Programme starts:		
9:00-9:10	Randi (DSBS) Jonathan (FMS)	Welcome
Session 1:		**Topics in group sequential designs**
9:10 – 9:45	Corine Baayen (Ferring)	Design and analysis of group sequential trials for repeated measurements when pipeline data occurs: a comparison of methods
9:45 – 10:20	Henrik Thomsen (Novo)	Family-wise error for multiple time-to-event endpoints in a group sequential design
Break **(30 minutes)**		
Session 2:		**Working as a pharmaceutical statistician**
10:50 – 12:20	Anna Berglind (Novo) Jonas Häggström (Cytel) Niklas Berglind (AstraZeneca)	Medical statistics in practice – different ways of making a difference
Lunch **12:20-13:30**		

Timeslot	Speaker	Title
Session 3:		**Utilization of historical data**
13:30 – 14:05	Martin Bøg (Novo)	Historical Borrowing
14:05-14:40	Daniel Jonker (Ferring)	Advancing Precision Medicine with Innovative In Silico Approaches in Reproductive Medicine
Break **(20 minutes)**		
Session 4:		**Next Generation of young statisticians**
15:00-15:35 **On Teams**	Emilie Højbjerre-Frandsen (Novo & AAU, Ph.d. Berkeley US)	Prognostic Score Adjustment
15:35-16:00	Wrap up	
16:00		End of the day

Session 1: **Topics in group sequential designs**

Session lead: Carl-Fredrik

Design and analysis of group sequential trials for repeated measurements when pipeline data occurs: a comparison of methods

Corine Baayen

Talk based on the (submitted) tutorial paper:

Design and analysis of group sequential trials for repeated measurements when pipeline data occurs: a tutorial

Corine Baayen[1,4], Paul Blanche[2], Christopher Jennison[5], Brice Ozenne[2,3]

R-code available on Github: **DelayedGSD** package.

[1]Biometric Division, H. Lundbeck A/S, Valby, Denmark

[2]Department of Public Health, Section of Biostatistics, University of Copenhagen, Copenhagen, Denmark

[3]Neurobiology Research Unit and BrainDrugs, Copenhagen University Hospital, Rigshospitalet, Copenhagen, Denmark

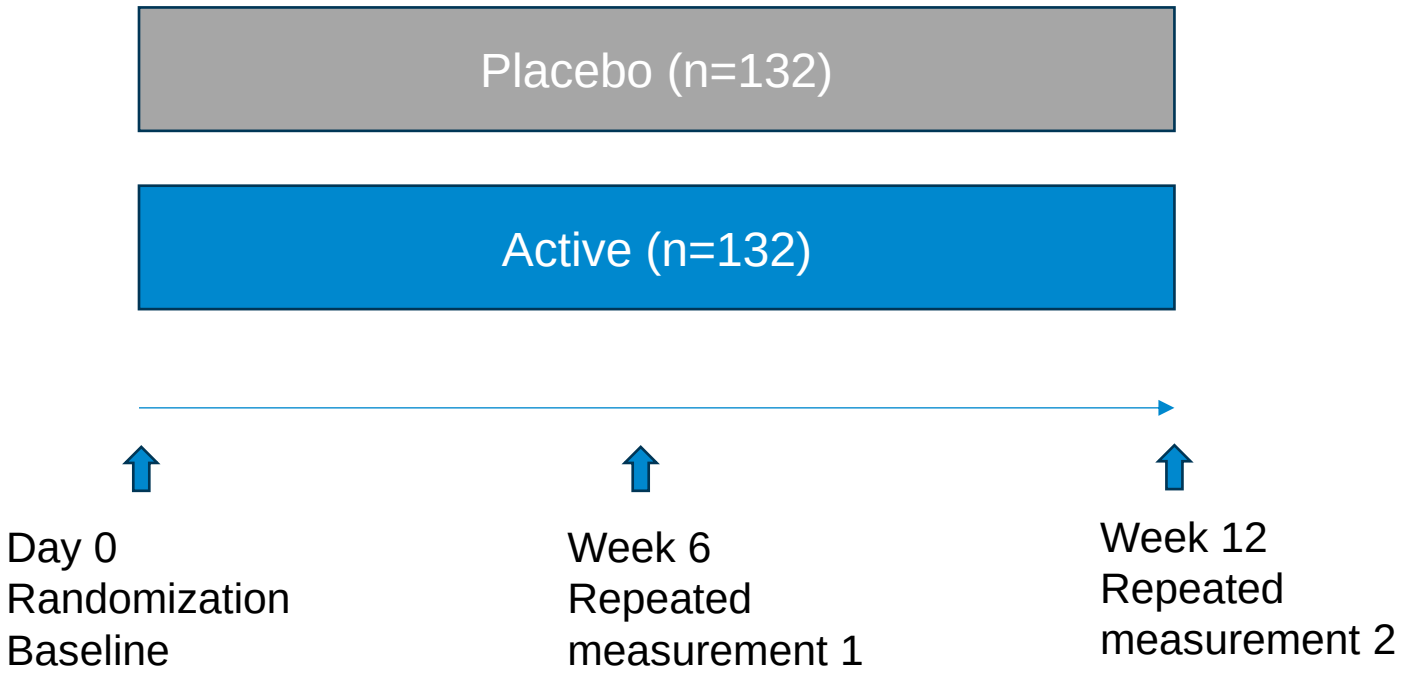
[4]Global Biometrics, Ferring Pharmaceuticals, Copenhagen, Denmark

[5]Department of Mathematical Sciences, University of Bath, Bath, United Kingdom

FERRING PHARMACEUTICALS

Motivating example: a phase 3 trial

Population: patients with a certain neurodegenerative disease
Primary endpoint: change from baseline to Week 12 on a continuous score
Sample size assumptions: power of 90%, one-sided significance level 2.5%, effect 1, sd 2.5, dropout 10%



Analysis: MMRM

Effect of interest θ: difference in change from baseline to Week 12 between the placebo and active arm

MMRM = Mixed model for repeated measurements

FERRING PHARMACEUTICALS

An interim analysis is included

Purpose:

- Stop early for efficacy
- Stop early for futility

Motivation

- Plan for a conservative effect size, but stop early in case effect is larger
- Quicker decision-making and lower average sample size

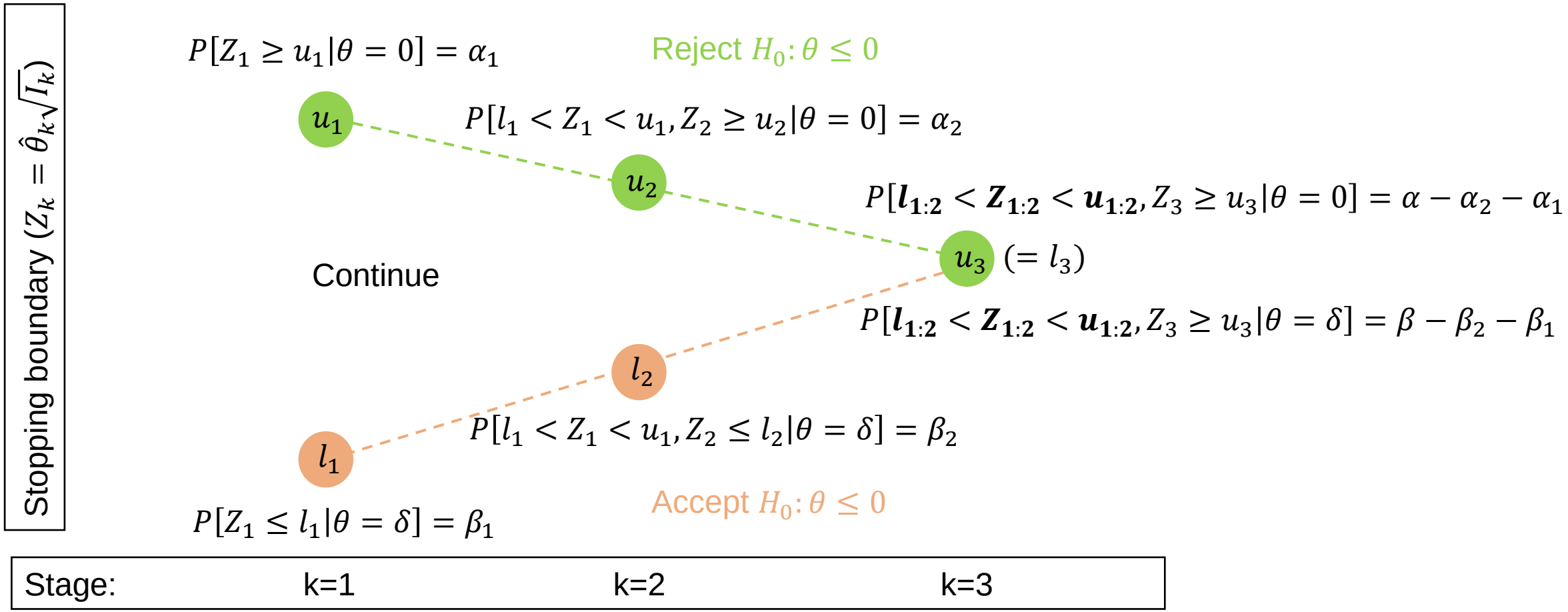
Timing:

When half (132) of the patients has either completed the 12 week treatment period or dropped out.

FERRING PHARMACEUTICALS

Stopping boundaries for a standard GSD (3 stage example)

Choose boundaries that control type I error level at α and type II error level at β



GSD = Group Sequential Design, $I_k = 1/\mathrm{var}(\hat{\theta}_k)$ is the observed information at stage *k*

The joint distribution of the test statistics

The cannonical distribution

To calculate the probability of a type I and type II error at each stage, e.g.

$$P[l_1 < Z_1 < u_1, Z_2 \geq u_2 | \theta = 0]$$

we need to understand the joint distribution of the test statistics Z_k

The observed information for θ at stage *k* equals $I_k = 1/\mathrm{var}(\hat{\theta}_k)$

For most common test statistics, the joint distribution is known and is called the cannonical joint distribution:

$$Z_k \sim N\left(\theta, \sqrt{I_k}\right)$$

$$Cov(Z_l, Z_{l'}) = \sqrt{I_l / I_{l'}}$$

FERRING PHARMACEUTICALS

Deciding on error levels spent at each stage

Error spending functions

We need to pre-define the type I and II error levels (α_k and β_k) that we wish to spend at each stage

A flexible approach to doing so is by using error spending functions

Error spending functions map the observed information at a stage to a cummulative error level to be spent by that stage. For example Kim and DeMets proposed:

$$\text{For the type I error: } f(I) = \alpha \min(1, \left\{\frac{I}{I_{max}}\right\}^{\rho})$$

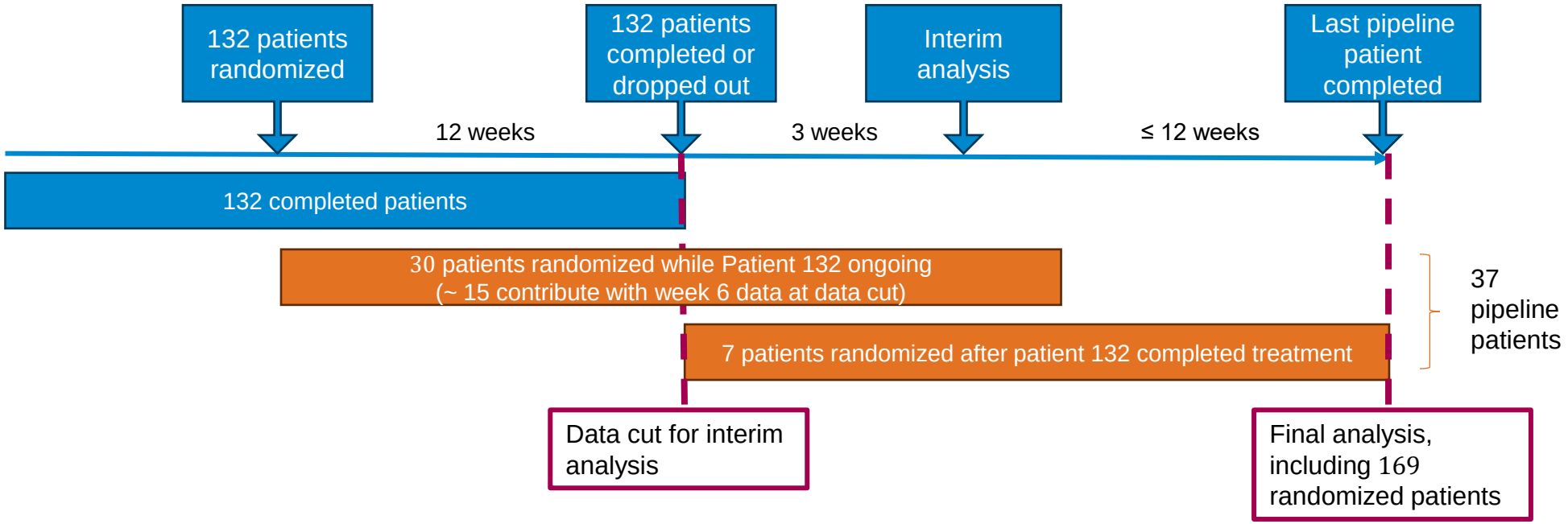
$$\text{For the type II error: } g(I) = \beta \min(1, \left\{\frac{I}{I_{max}}\right\}^{\rho})$$

Kim K, DeMets DL. Design and analysis of group sequential tests based on the type I error spending rate function. *Biometrika.* 1987;74(1):149–154.

FERRING PHARMACEUTICALS

Pipeline data (recruitment rate 2.5 patients/week)

At the time of the interim analysis, not all randomized patients will have completed the study



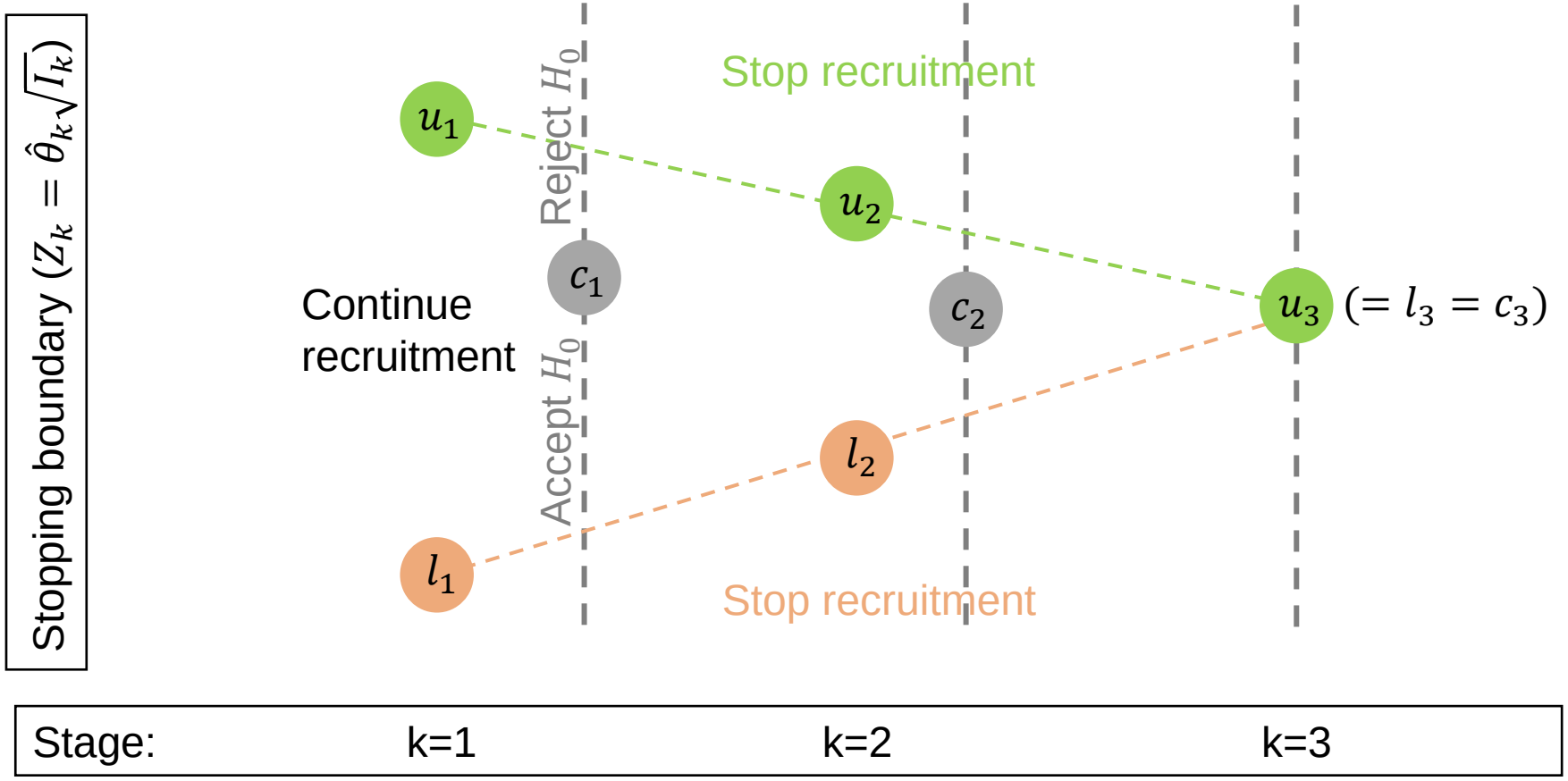
- Include the data from pipeline patients in the final analysis to achieve maximal precision.

GSD with pipeline data (3 stage example)

Hampson & Jennison, JRSS-B 75(1):3–54, 2013

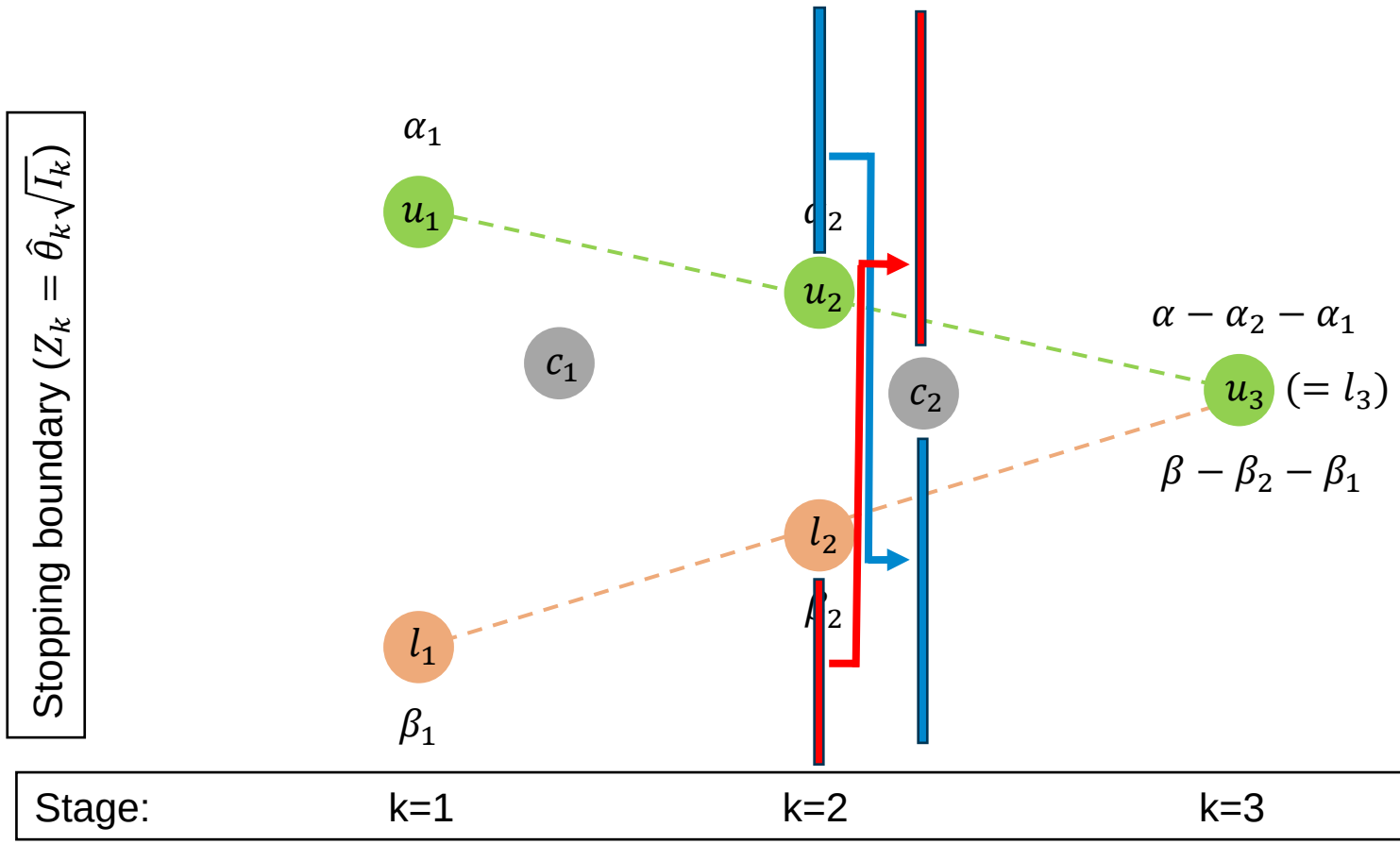
At interim: decide whether to stop recruitment based on all available data

At decision: after following up on all patients, final decision on whether to reject the null hypothesis



Decision boundaries – Method 1

Interim stopping boundaries u_k and l_k derived as for standard GSD (`type I and II error´ controlled at interim)



Calculate c_k by solving:

$$P\left[\boldsymbol{Z}_{\boldsymbol{1:(k-1)}} \in \boldsymbol{C}_{\boldsymbol{1:(k-1)}}, Z_k \geq u_k, \tilde{Z}_k < c_k | \theta = 0\right]$$
$$=$$
$$P\left[\boldsymbol{Z}_{\boldsymbol{1:(k-1)}} \in \boldsymbol{C}_{\boldsymbol{1:(k-1)}}, Z_k \leq l_k, \tilde{Z}_k \geq c_k | \theta = 0\right],$$

with

$$C_k = (l_k, u_k)$$

and

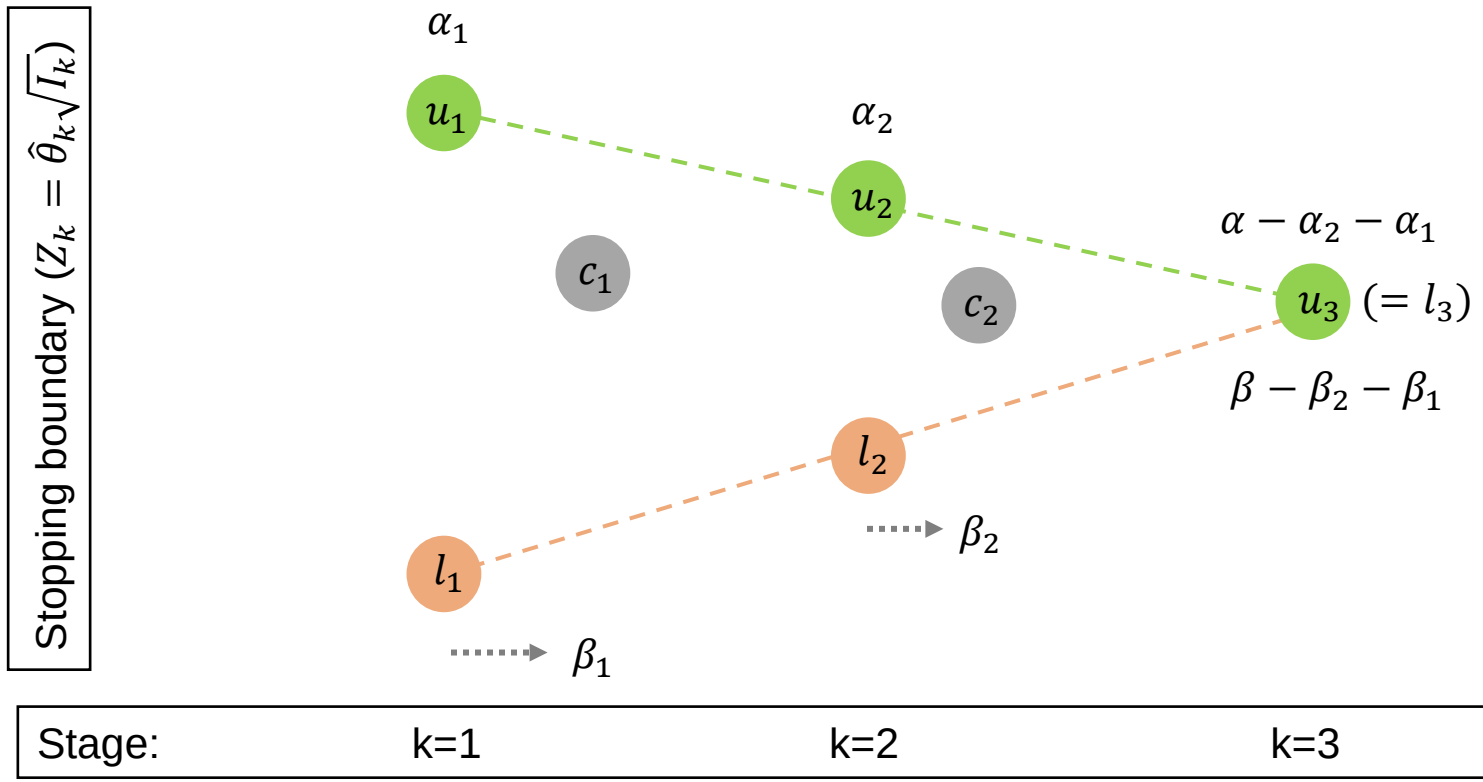
$$\tilde{Z}_k = \tilde{\theta}_k\sqrt{\tilde{I}_k}$$

the Z-statistic at the decision analysis.

Note that the type II error spent at decision analysis k is $< \beta_k$, leading to an increase in power, as there is more data than at the interim analysis.

Decision boundaries – Method 2

Same principle, but aims to achieve planned power exactly



At **interim** analys k:

1. Calculate u_k as usual
2. Predict the information at the decision analysis: $\tilde{I}_k$
3. Simultaneously search for l_k and c_k such that the type II error spent at decision equals β_k (c_k defined as before)

At **decision** analys k: recalculate c_k

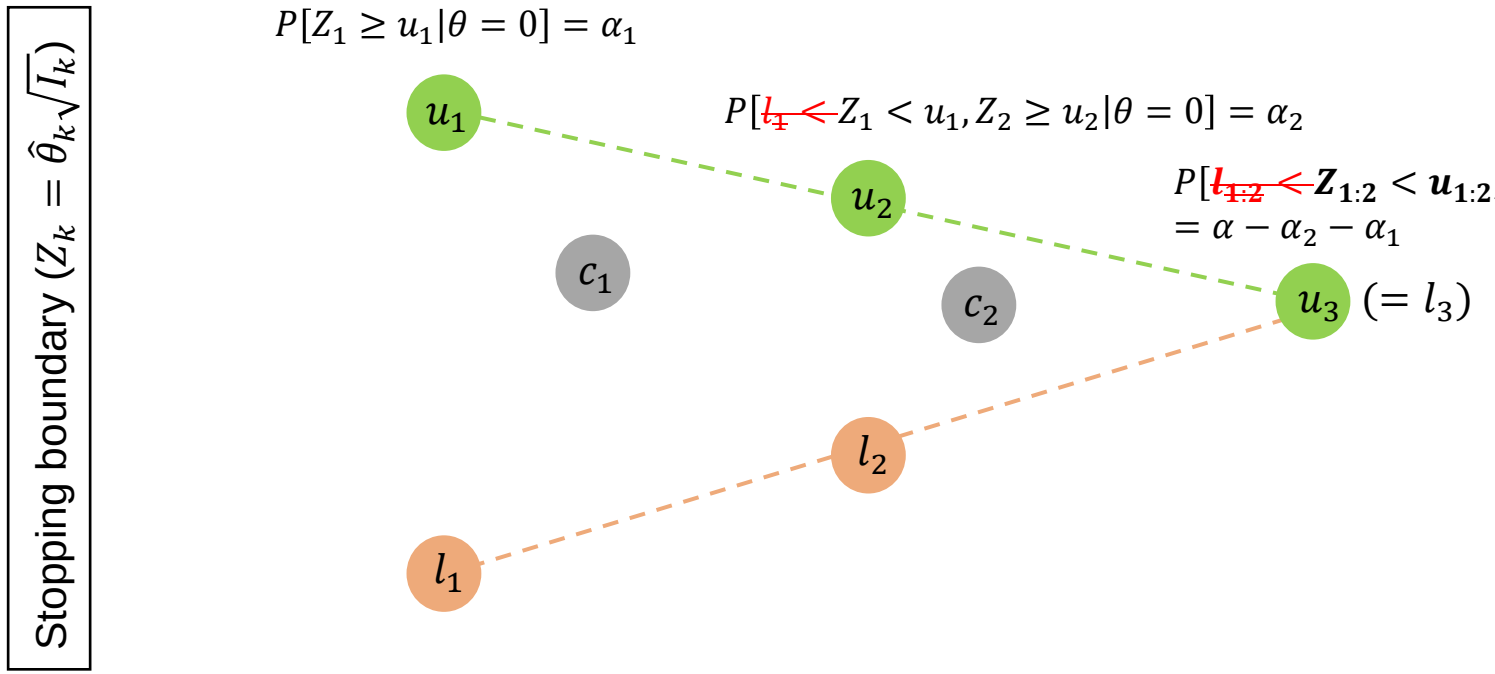
Power likely closer to planned, but may not match exactly if $\tilde{I}_k$ was not correctly predicted.

FERRING PHARMACEUTICALS

A proposal for a non-binding futility rule for Methods 1 and 2

Boundaries should maintain type I error control even if the futility rule is ignored

Boundaries at the interim analyses similar as for a standard GSD with non-binding futility boundaries



Calculate c_k by solving:

$$P\left[\boldsymbol{Z_{1:(k-1)}} < \boldsymbol{u_{1:(k-1)}}, Z_k \geq u_k, \tilde{Z}_k < c_k|\theta = 0\right] =$$

$$P\left[\boldsymbol{Z_{1:(k-1)}} < \boldsymbol{u_{1:(k-1)}}, Z_k \leq l_k, \tilde{Z}_k \geq c_k|\theta = 0\right],$$

Stage:	k=1	k=2	k=3

Proof provided in Baayen et al. (submitted)

Choice of critical value at the decision analysis

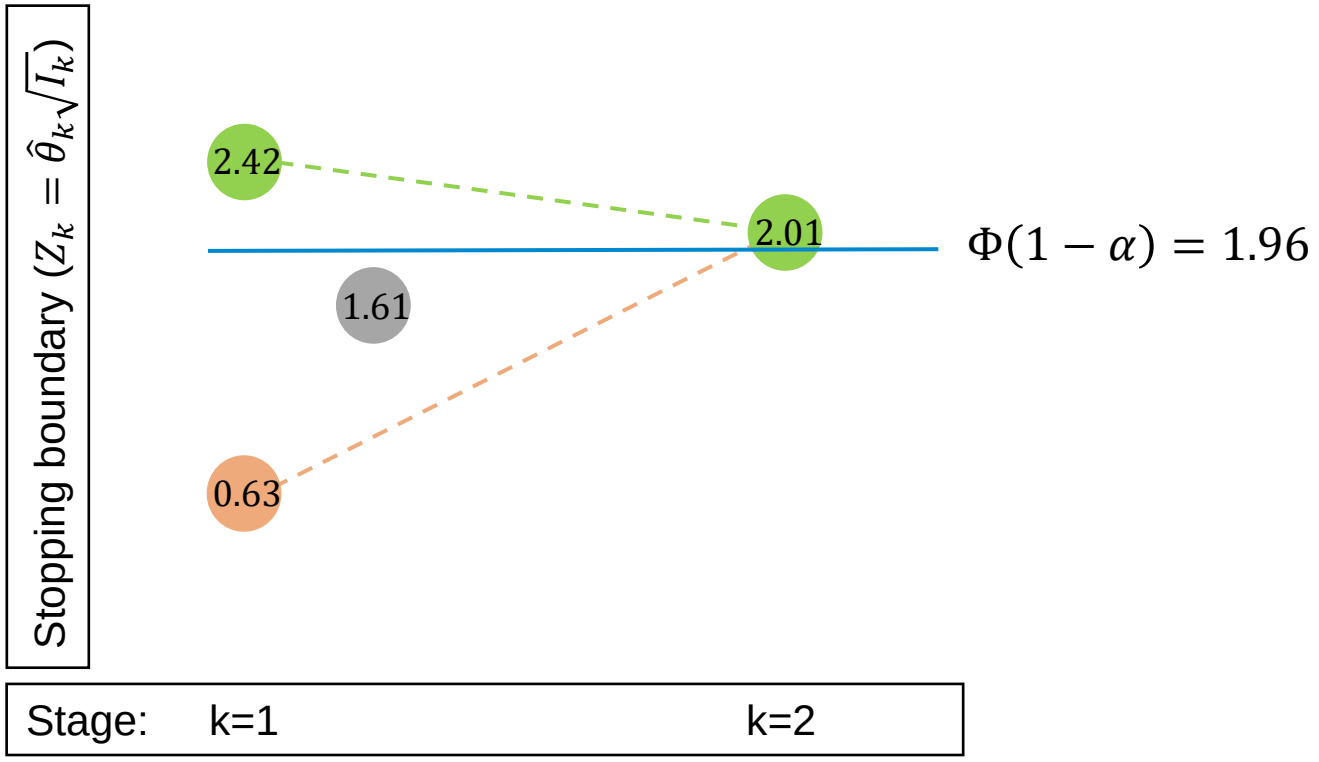
c_k can be substantially lower than the critical value for a fixed design $\Phi(1-\alpha)$

For reasons of credibility we might suggest $c'_k = \max(c_k, \Phi(1-\alpha))$

This ensures that concluding efficacy will never be easier than if we had obtained the same data without an interim analysis (fixed trial)

Conservative type I error control

May reduce power



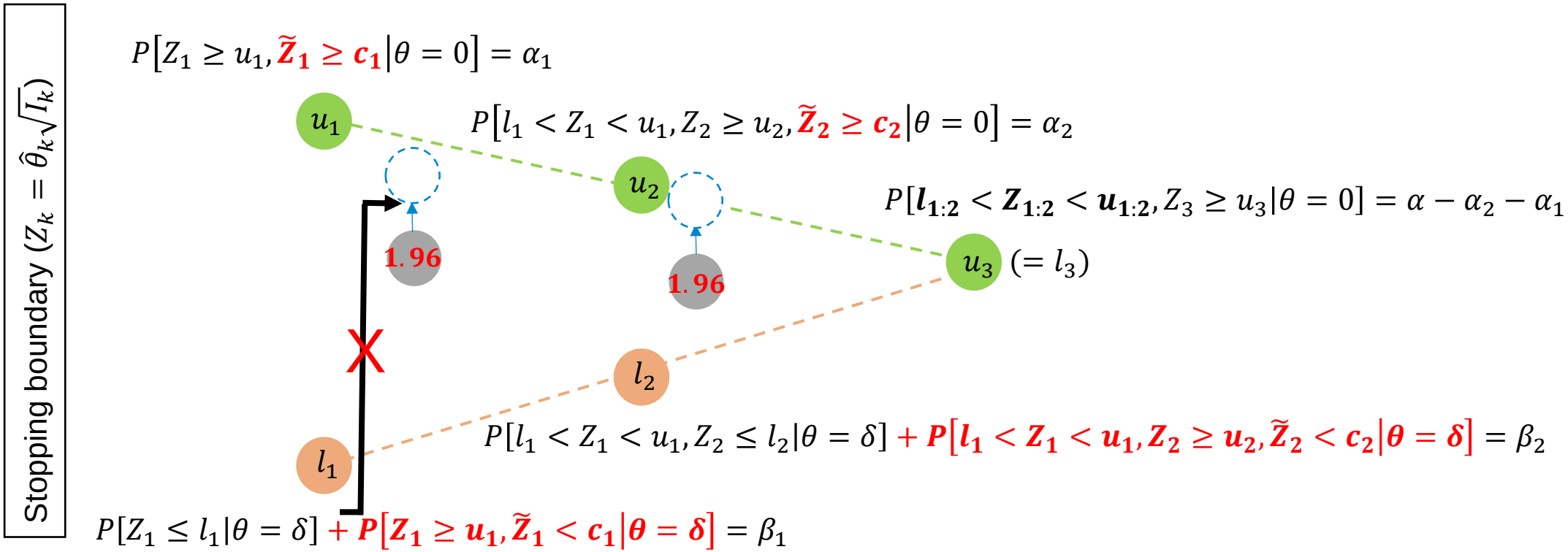
Planned boundaries for the case study

Method 3 (Jennison 2022, course slides)

Avoiding reversals from futility to efficacy and anticipating $c_k \geq \Phi(1-\alpha)$

If at interim recruitment is stopped due to negative results ($Z_k < l_k$):

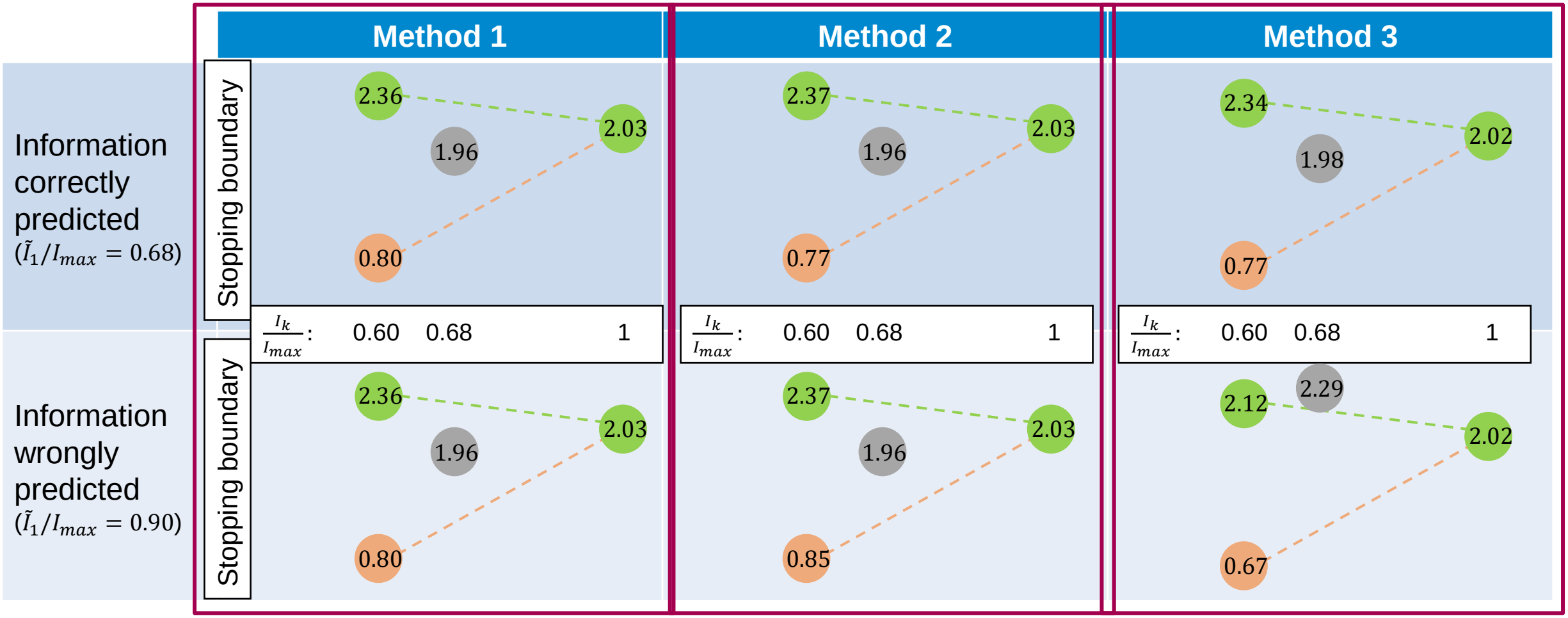
- One may wish to stop following/treating patients for ethical and practical reasons
- Reversals to a positive result at decision ($\tilde{Z}_k \geq c_k$) may not be considered credible



Stage:	k=1	k=2	k=3

Boundaries for the case study according to all methods

Similar boundaries at planning stage, results may differ (somewhat) when executing



Same, no dependence on future info

With 1 interim, only futility bnd differ

All boundaries depend on future info

Comparison of methods (simulations show only minor differences)

	Method 1		Method 2		Method 3
	No constraint	$c_k \geq \Phi(1-\alpha)$	No constraint	$c_k \geq \Phi(1-\alpha)$	$c_k \geq \Phi(1-\alpha)$
Type I error			**Controlled**		
Power	Can be higher (~0.3% point)	Can be lower (~0.5% point)	Yes (if correct $\tilde{I}_k$ prediction)	Can be lower (~0.5% point)	Yes (if correct $\tilde{I}_k$ prediction)
Information extrapolation	**No**	**No**	**Yes**	**Yes**	**Yes**
Non-binding futility rule			**Optional**		
Reversal fut to eff	**Yes** (0.05-0.5%)	**Yes** (0-0.25%)	**Yes** (0.05-0.5%)	**Yes** (0-0.25%)	**No**
Correct inference	**Yes** (Median Unbiased Estimate, confidence interval and p-value), **but** see presentation at DSTS 2-day meeting and paper Baayen et al. for details				

Interesting approach when information at decision difficult to predict (e.g. delayed endpoints without short term measurements)?

Preferred choice for case study. Generally preferrable if information can be predicted with reasonable accuracy?

Different approaches to handling pipeline data

Inclusion of pipeline data
Incorporate in hypothesis test (primary analysis upon observing all pipeline data)
Do not incorporate in hypothesis test (primary analysis at interim, ignoring pipeline data)

Asikanius, E., et al., 2024. Considerations for the planning , conduct and reporting of clinical trials with interim analyses. *ArXiv:* [2410.01478v1] Considerations for the planning, conduct and reporting of clinical trials with interim analyses

Schüürhuis, S., et al., 2024. A two-stage group-sequential design for delayed treatment responses with the possibility of trial restart. *Statistics in Medicine*, *June 2023*, 1–21. https://doi.org/10.1002/sim.10061

FERRING PHARMACEUTICALS

Disclaimer:
Views and opinions expressed are those of the speaker and not necessarily Novo Nordisk

Received: 16 August 2023 | Revised: 15 May 2024 | Accepted: 22 May 2024
DOI: 10.1002/sim.10132

RESEARCH ARTICLE

Statistics in Medicine WILEY

Familywise error for multiple time-to-event endpoints in a group sequential design

Henrik F. Thomsen[1] | Nanna L. Lausvig[2] | Christian B. Pipper[2,3] | Søren Andersen[2] | Lars H. Damgaard[2] | Scott S. Emerson[4] | Henrik Ravn[2]

[1]Department of Biostatistics, Novo Nordisk A/S, Aalborg, Denmark
[2]Department of Biostatistics, Novo Nordisk A/S, Søborg, Denmark
[3]Department of Public Health, University of Southern Denmark, Odense, Denmark
[4]Department of Biostatistics, University of Washington, Seattle, Washington,

Correspondence
Henrik F. Thomsen, Novo Nordisk A/S, Alfred Nobels Vej 27, DK-9220 Aalborg Øst, Denmark.
Email: hfth@novonordisk.com

We investigate the familywise error rate (FWER) for time-to-event endpoints evaluated using a group sequential design with a hierarchical testing procedure for secondary endpoints. We show that, in this setup, the correlation between the log-rank test statistics at interim and at end of study is not congruent with the canonical correlation derived for normal-distributed endpoints. We show, both theoretically and by simulation, that the correlation also depends on the level of censoring, the hazard rates of the endpoints, and the hazard ratio. To optimize operating characteristics in this complex scenario, we propose a simulation-based method to assess the FWER which, better than the alpha-spending approach, can inform the choice of critical values for testing secondary endpoints.

KEYWORDS
familywise error rate, group sequential design, secondary endpoints, time to event endpoints

1 | INTRODUCTION

Large event-driven clinical trials often employ a group sequential design (GSD) with a single planned interim analysis of a primary time-to-event (TTE) endpoint to allow for early stopping for efficacy or futility. Typically, such trials are designed using a parallel-group 1:1 randomization to active treatment and comparator treatment and encompass a primary and at least one confirmatory secondary endpoint, which are tested using a stage-wise hierarchical strategy. That is, the secondary endpoint(s) are tested only if the primary and preceding secondary endpoints are statistically significant. In this setup, the secondary endpoints are tested at most once; either at the interim analysis, the final analysis or not at all. Examples of the use of this hierarchical test strategy are abundant.[1-3]

The method evaluated in the present study is inspired by the motivating example provided later in this article (Section 4). In summary, we construct a plausible scenario for a trial with a primary composite TTE endpoint comprising the cardiovascular event types non-fatal myocardial infarction (MI), non-fatal stroke, and cardiovascular death. In addition, the scenario trial includes a single-component secondary endpoint consisting of cardiovascular death. Of note, the primary endpoint includes the secondary endpoint plus a second component of non-fatal MI or non-fatal stroke. Accordingly, concordance exists between the primary and secondary endpoints, which contributes to their correlation. Additional correlation originates from the inherent correlation between the endpoint components. These two sources of correlation of the endpoints are investigated individually and in conjunction to assess how they influence the correlation of the test statistics and the familywise error rate (FWER).

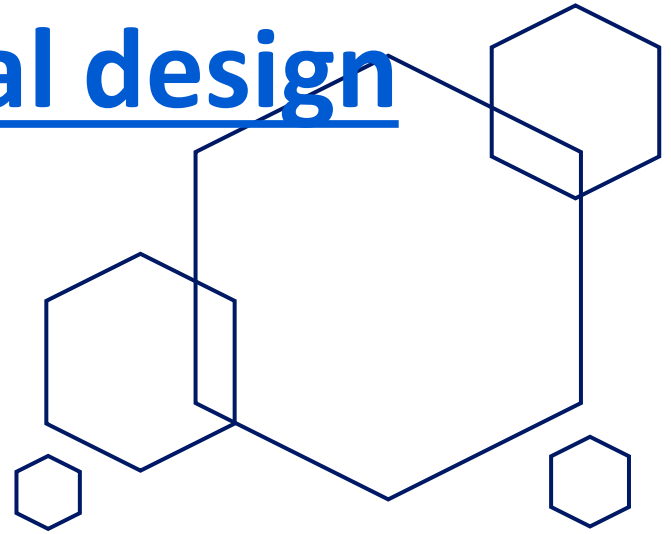
Controlling the overall FWER is crucial, not least in the design of trials supporting regulatory decision making. Thus, ways to optimize the statistical power while controlling the FWER are critically needed and even minor enhancements of

novo nordisk®

Familywise error for multiple time-to-event endpoints in a group sequential design

Henrik F. Thomsen

05NOV2024



Examples of endpoints in outcome trials

Primary endpoint

- MACE

Confirmatory secondary

- CV-death
- Composite heart failure
- All-cause death

Primary endpoint

- Composite CKD

Confirmatory secondary

- eGFR slope
- MACE
- All-cause death

Primary endpoint

- MACE

Confirmatory secondary

- Composite CKD
- CV-death
- MALE (major adverse limb events)

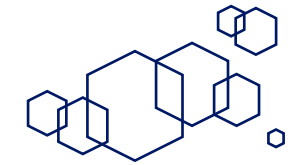
CV, cardiovascular; CKD, chronic kidney disease

Setting the stage

- **Group sequential design (GSD)**: A design where a hypothesis is tested multiple times based on an increasing amount of data.
- **Alpha-spending**: Method for handling the multiplicity-issues of testing the same hypothesis multiple times – utilizes the fact that the hypotheses are strongly correlated.
- **Hierachical testing**: Method for handling multiplicity-issues when testing more than one hypothesis.

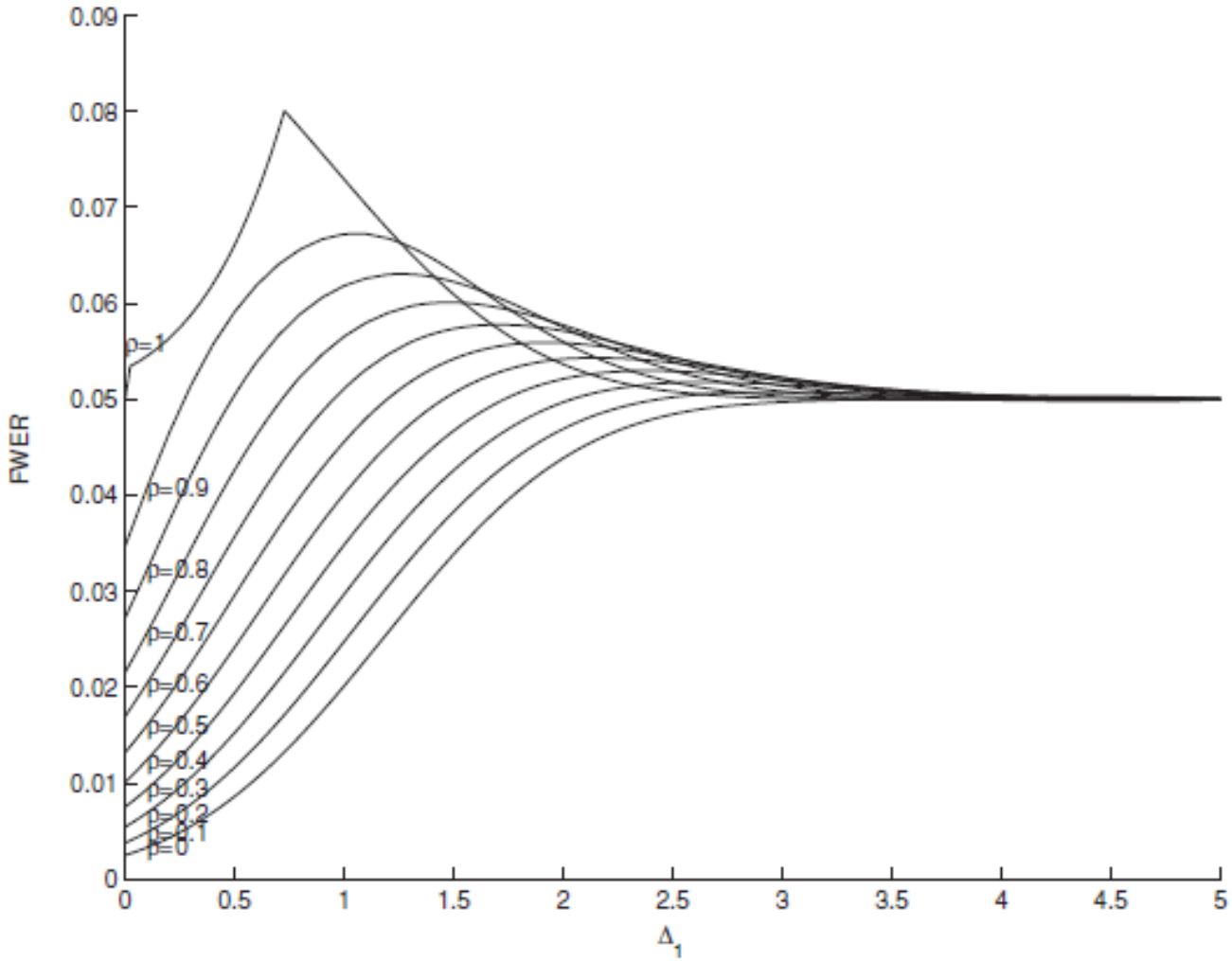
$$\begin{pmatrix} Z_{p,i} \\ Z_{s,i} \\ Z_{p,f} \\ Z_{s,f} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \sqrt{t}\delta_p \\ \sqrt{t}\delta_s \\ \delta_p \\ \delta_s \end{pmatrix}, \begin{bmatrix} 1 & \rho_E & \sqrt{t} & \rho_E\sqrt{t} \\ \cdot & 1 & \rho_E\sqrt{t} & \sqrt{t} \\ \cdot & \cdot & 1 & \rho_E \\ \cdot & \cdot & \cdot & 1 \end{bmatrix}\right), \tag{1}$$

$$P(Z_{p,i} \geq C_{p,i}, Z_{s,i} \geq C_{s,i}) + P(Z_{p,i} < C_{p,i}, Z_{p,f} \geq C_{p,f}, Z_{s,f} \geq C_{s,f}), \tag{2}$$



Testing a Primary and a Secondary Endpoint in a Group Sequential Design

Ajit C. Tamhane,[1,*] Cyrus R. Mehta[2,**] and Lingyun Liu[3,***]



Added complications

Primary endpoint is MACE: time to first of non-fatal MI/stroke and CV-death

Secondary endpoint: time to CV-death

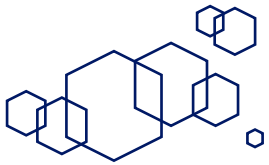
- 2 types of correlation.
  - ‘ordinary’ correlation of components
  - Concordance
- There is (substantial) censoring of the endpoints.

Simulation setup (base case)

- 1:1 randomization ratio.
- Comparator hazard rate for primary endpoint 0.0425.
- Hazard ratio for primary endpoint 0.85.
- One-sided significance level of 2.5%.
- Power for the primary endpoint of 90%.
- One interim analysis at 2⁄3 of the planned information.
- Resulting in the need to accrue 1610 events at final analysis, and 1074 events at interim analysis (O'Brien-Fleming alpha-spending for primary).
- An accrual time of 0.5 years, and a total duration of 5 years.
- Assume the event-times are following an exponential distribution.
- Only administrative censoring.
- Calculate the 4 Z-stats for each sim.
  - Based on these calculate the var matrix of (1) and then use (2) to calculate the FWER.

‘Ordinary’ correlation – no concordance

- Simulate the time-to-events by means of a normal copula: N -> uniform -> exponential dist.
- How do the endpoint correlation translate over into the Z-stats correlation.
  - For simplicity, we present only the observed correlation of the Z-statistics for the test of the primary endpoint and secondary endpoint at the **final** analysis.
  - The rates are given as λ_p and λ_s for the comparator arm, and $HR_p\lambda_p$ and $HR_s\lambda_s = \lambda_s$ as HR_s= 1.



Conclusions

- Higher censoring leads to lower Z-stat correlation (attenuation)
- Higher endpoint correlation leads to higher Z-stat correlation
- The distribution between primary and secondary rates plays a role
- No clear effect for different effect-sizes of the primary endpoint

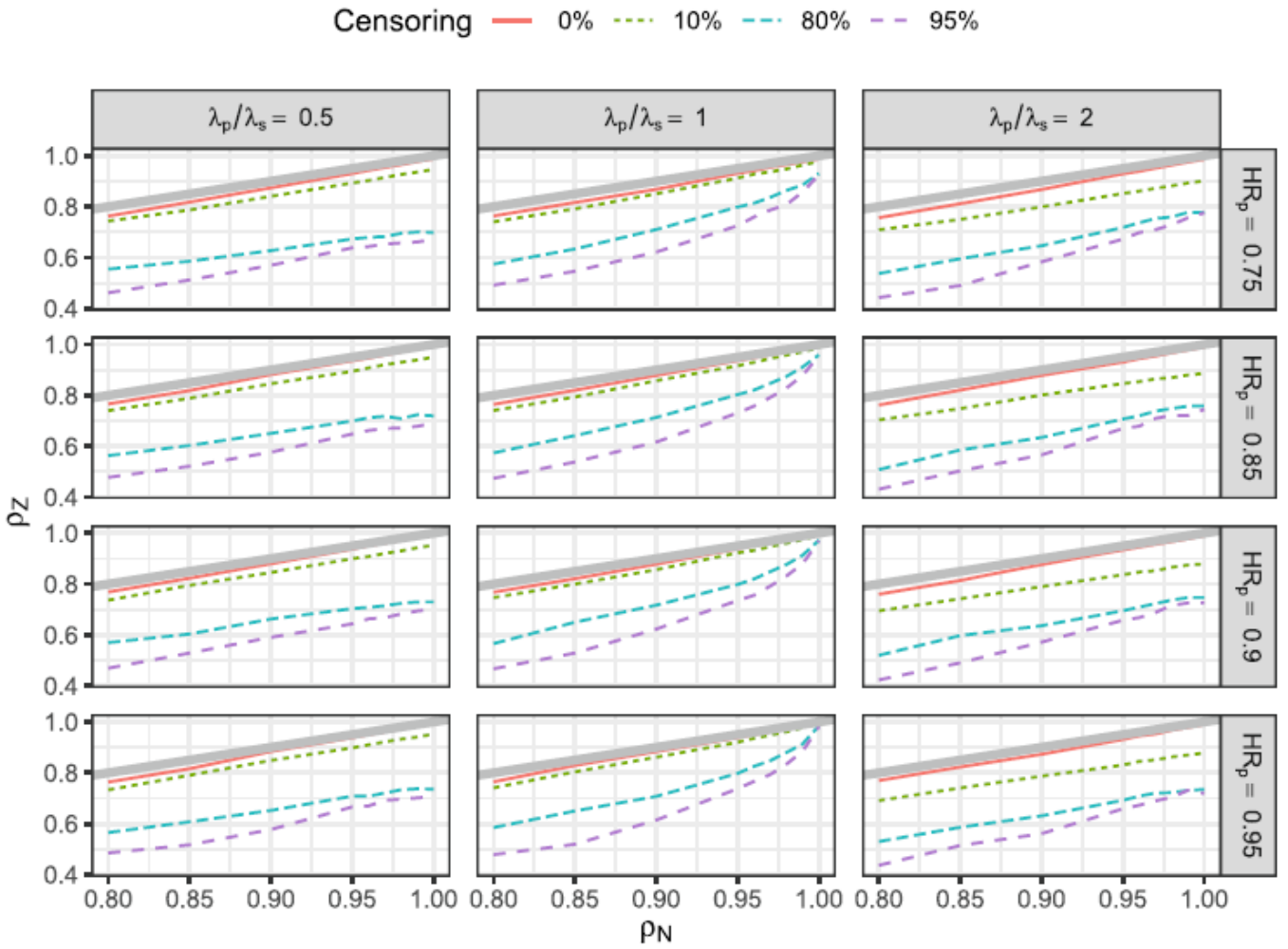


FIGURE 1 ρ_Z as a function of ρ_N, HR_p, λ_p/λ_s, and level of censoring at final analysis. The wide gray line is representing $\rho_Z = \rho_N$.

Conclusions

- Highest FWER for correlation of 1 (like in Tamhane)
- Very little or no inflation for HR <= 0.90
- Highest inflation seen for equal rates for primary and secondary endpoint
- If correlation is below say 0.9 no inflation
- If rate of primary is twice that of the secondary we see no inflation

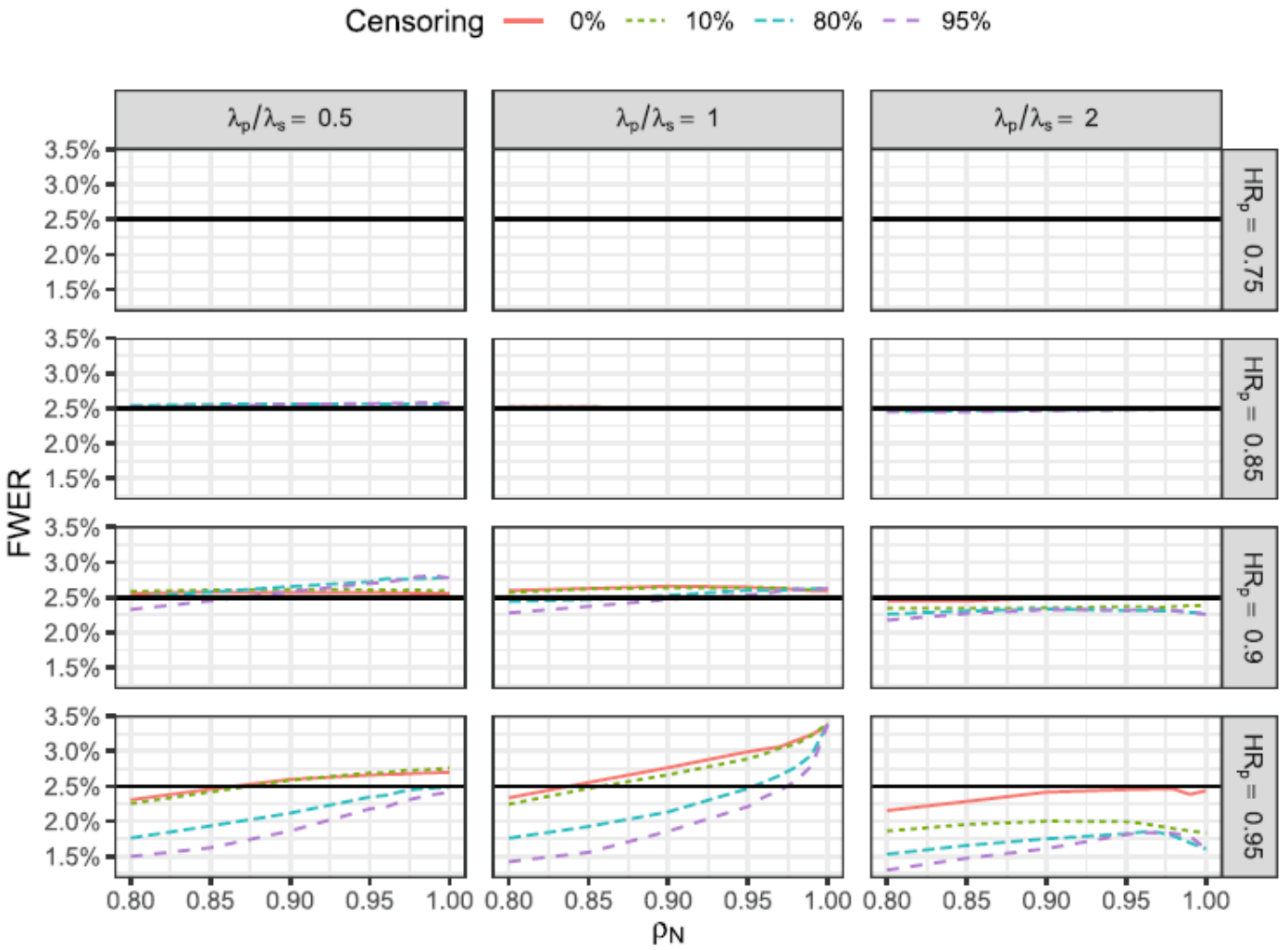
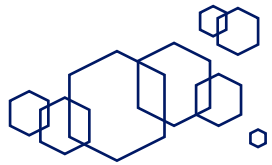


FIGURE 2 FWER for as a function of ρ_N, HR_p, λ_p/λ_s, and level of censoring at final analysis.

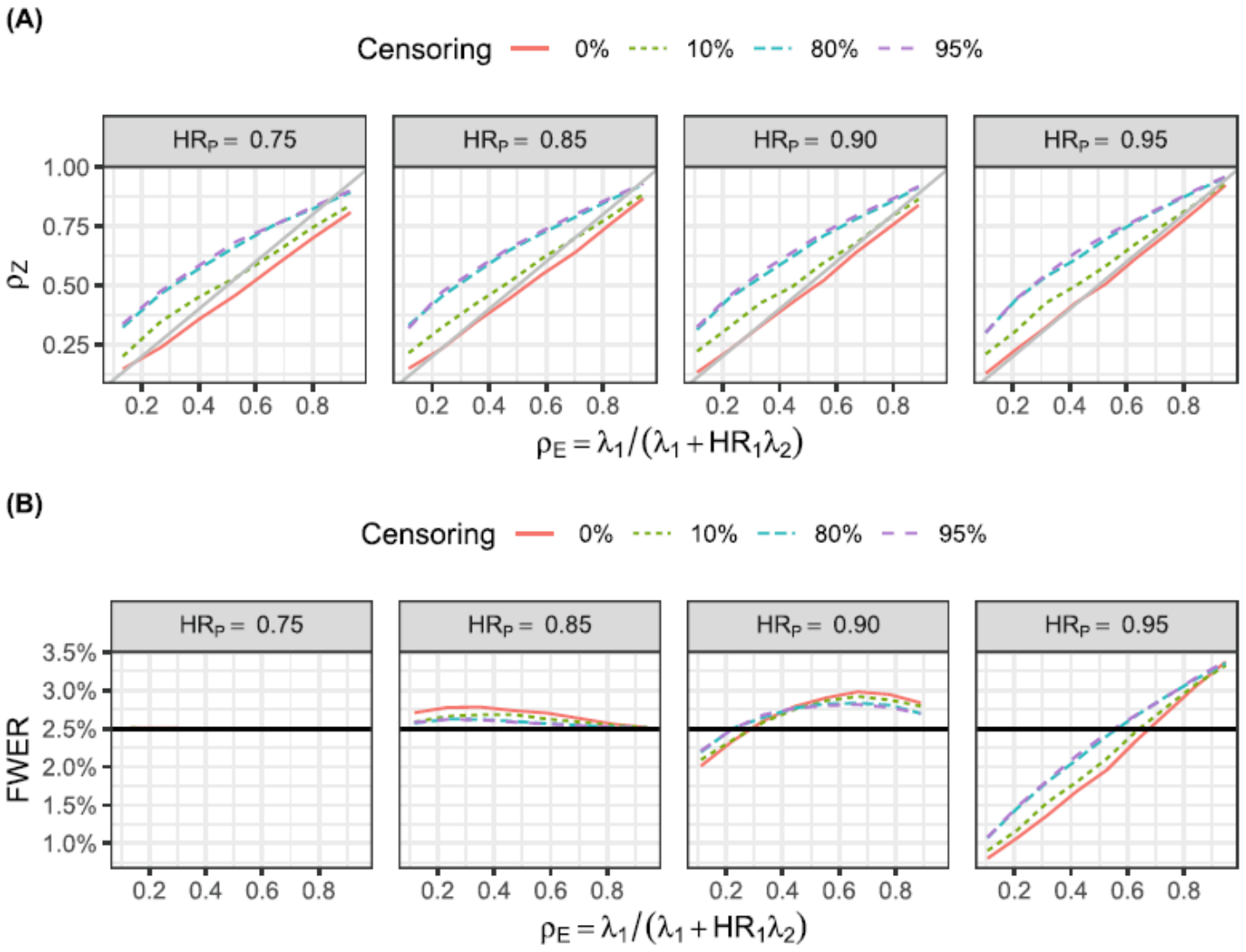
Concordance

- Simulate two **independent** components from an exponential distribution
  1. Non-fatal MI/Stroke
  2. CV-death
- Primary endpoint is the minimum of the time-to-event for the two components
  - Due to the independence, this endpoint is also exponentially distributed
- The secondary endpoint is component 2 (CV-death)



Conclusions

- A higher degree of censoring leads to a higher correlation between the Z-statistics.
- When the endpoint correlation increases, the correlation of the Z-statistics increases.
- The correlation of the Z-statistics is in general higher than the endpoint correlation.
- Finally, we show that marked inflation of the FWER occurs only when the effect size for the primary endpoint is small and the correlation (concordance) between endpoints pronounced.



Combined

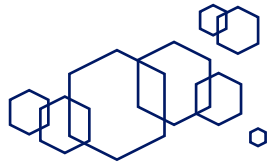
- The primary endpoint encompasses non-fatal myocardial infarction (MI), non-fatal stroke, and cardiovascular death, while the secondary endpoint solely consists of cardiovascular death. The two components—non-fatal MI/stroke and cardiovascular death—will be correlated
- Interim and final analyses are conducted after accruing 1074 and 1610 events, respectively, and the trial enroll 4729 subjects in each treatment arm
- For the analysis of the primary endpoint, we use O'Brien-Fleming alpha-spending critical values of 2.51 and 1.99 for interim the and final analysis, respectively
- Things get a little more complicated with correlated components (eg the primary endpoint is no longer exponentially distributed due to the components being correlated)
- We use NN data to get relevant estimates rates and correlations

Estimates from historic data

Placebo hazard rates for cardiovascular death are set to 0.014 and confined to the interval [0.005; 0.02]

Placebo hazard rates for non-fatal myocardial infarction/non-fatal stroke are set to be 0.03 and confined to [0.02; 0.04].

The Spearman correlation is set to 0.4 confined to the interval [0; 0.7]



Worst case scenario

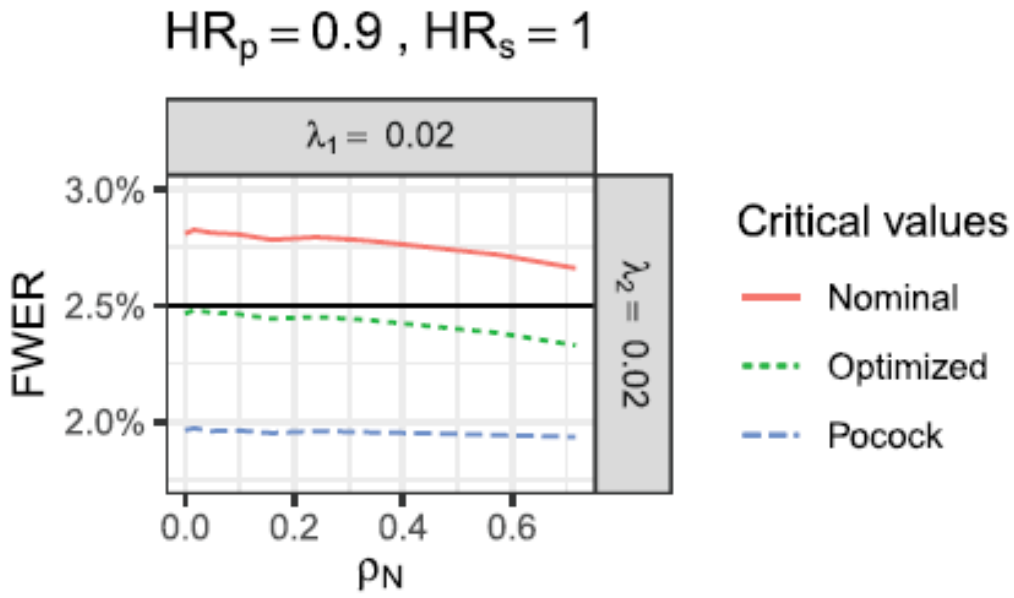


FIGURE 4 FWER as a function of ρ_N, using nominal (1.96), Pocock (2.07, 2.25), and optimized (2.015) critical values for the secondary endpoint, for the worst case combination of λ_1, λ_2 and HR_p.

Power gain

Using the optimized critical values, instead of the alpha-spending Pocock approach in the “base case” (HR_P = 0.85, HR_S = 0.85, placebo hazard rate for the primary endpoint of 0.0425, placebo hazard rate for the secondary endpoint of 0.0125, and a correlation between components of 0.4), we show an increase in the power of the secondary endpoint (**from 29.4% to 33.5%**).

Conclusions

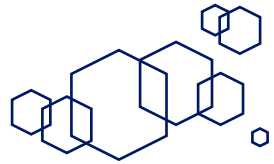
In general, we observe that a high between Z-statistics correlation is associated with a high FWER.

The correlation between the Z-statistics is attenuated by censoring of the endpoints when the endpoints are correlated without sharing a component, and conversely when the correlation between the endpoints is solely from concordance.

The correlation of the Z-statistics further depends both on the hazard rates of the endpoints and the hazard ratios.

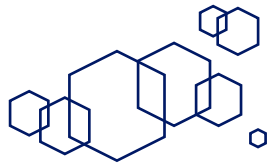
We have given tools to evaluate the FWER by means of simulations and suggestions on how to optimize critical values under FWER control in a GSD utilizing a hierarchical testing strategy.

The proposed use of critical values from alpha-spending GSD will in some real world scenarios be unnecessarily restrictive.



Limitations

- Both the primary and secondary endpoint are TTE endpoints, but the methods are straightforward to apply with endpoints with other distributions.
- Only one secondary endpoint.
- The primary endpoint consists of 2 components, and the secondary consists of one of those components.
- A normal copula is used when generating correlated exponential distributed data.
- We are only looking at differing levels of administrative censoring.
- We are using an exponential distribution for the TTE endpoints.



Break

30 minutes

Today's programme

Timeslot	Speaker	Title
8:30 – 9:00		Breakfast and arrival
Programme starts:		
9:00-9:10	Randi (DSBS) Jonathan (FMS)	Welcome
Session 1:		**Topics in group sequential designs**
9:10 – 9:45	Corine Baayen (Ferring)	Design and analysis of group sequential trials for repeated measurements when pipeline data occurs: a comparison of methods
9:45 – 10:20	Henrik Thomsen (Novo)	Family-wise error for multiple time-to-event endpoints in a group sequential design
Break **(30 minutes)**		
Session 2:		**Working as a pharmaceutical statistician**
10:50 – 12:20	Anna Berglind (Novo) Jonas Häggström (Cytel) Niklas Berglind (AstraZeneca)	Medical statistics in practice – different ways of making a difference
Lunch **12:20-13:30**		

Timeslot	Speaker	Title
Session 3:		**Utilization of historical data**
13:30 – 14:05	Martin Bøg (Novo)	Historical Borrowing
14:05-14:40	Daniel Jonker (Ferring)	Advancing Precision Medicine with Innovative In Silico Approaches in Reproductive Medicine
Break **(20 minutes)**		
Session 4:		**Next Generation of young statisticians**
15:00-15:35 **On Teams**	Emilie Højbjerre-Frandsen (Novo & AAU, Ph.d. Berkeley US)	Prognostic Score Adjustment
15:35-16:00	Wrap up	
16:00		End of the day

Session 2: **Working as a pharmaceutical statistician**

Session lead: Mikala

Medical statistics in practice

different ways of making a difference

Anna Berglind, PhD
Vice President Biostatistics,
Rare Disease & Advanced Therapies
Novo Nordisk

Niklas Berglind, PhD
Global Product Leader
CV Renal and Metabolism
AstraZeneca

Jonas Häggström, PhD
Vice President
Global Health
Cytel

5 Nov 2024

Making a difference

How can statisticians make a real difference in medicine and impact patients and their families?

Pride in what we do

Examples of when you and/or your stats colleagues did something that really made a difference

Outside perspective

What are areas where statisticians are uniquely equipped to make a difference?
What to do and what not to do?

Statisticians' role in data science

Exciting times for broader data science area
Nobel prizes and enormous focus on AI
Risk or opportunity for statisticians?
Any advice for us as a skill?

Lunch

60 minutes

Today's programme

Timeslot	Speaker	Title
8:30 – 9:00		Breakfast and arrival
Programme starts:		
9:00-9:10	Randi (DSBS) Jonathan (FMS)	Welcome
Session 1:		**Topics in group sequential designs**
9:10 – 9:45	Corine Baayen (Ferring)	Design and analysis of group sequential trials for repeated measurements when pipeline data occurs: a comparison of methods
9:45 – 10:20	Henrik Thomsen (Novo)	Family-wise error for multiple time-to-event endpoints in a group sequential design
Break **(30 minutes)**		
Session 2:		**Working as a pharmaceutical statistician**
10:50 – 12:20	Anna Berglind (Novo) Jonas Häggström (Cytel) Niklas Berglind (AstraZeneca)	Medical statistics in practice – different ways of making a difference
Lunch **12:20-13:30**		

Timeslot	Speaker	Title
Session 3:		**Utilization of historical data**
13:30 – 14:05	Martin Bøg (Novo)	Historical Borrowing
14:05-14:40	Daniel Jonker (Ferring)	Advancing Precision Medicine with Innovative In Silico Approaches in Reproductive Medicine
Break **(20 minutes)**		
Session 4:		**Next Generation of young statisticians**
15:00-15:35 **On Teams**	Emilie Højbjerre-Frandsen (Novo & AAU, Ph.d. Berkeley US)	Prognostic Score Adjustment
15:35-16:00	Wrap up	
16:00		End of the day

Session 3: **Utilization of historical data**

Session lead: Kyle

Historical Borrowing feat. Bayesian Dynamic Borrowing

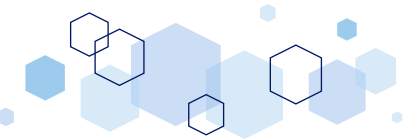
Joint FMS-DSBS Meeting

Martin Bøg (AXBQ), Biostatistics
HTA
Novo Nordisk

Agenda

- Introduction and rationale for borrowing
- Bayesian Dynamic Borrowing in a nutshell
- Elements of a BDB design
- A case example from Novo Nordisk
- Regulatory Outlook
- Summary

Disclaimer: Views expressed are my own, and do not necessarily represent Novo Nordisk



Why borrow?

Conclusion

All the forces in the world are not so powerful as an idea whose time has come.
—Victor Hugo[94]

There continues to be a sense of urgency in developing medicines for patients in need. Patients, academics, drug development companies, and regulators are all incentivized to accelerate our ability to test new interventions for efficacy and safety while minimizing subject exposure. Regulators have a record of accepting historical control data for interventions for medical devices and/or indications with small populations.

The methods covered in this paper give us the tools to use fewer subjects in late-phase confirmatory clinical trials. It is our opinion that this is an idea whose time has come. The industry and regulatory science has matured to the point where high-quality data exists to support these approaches; the statistical methods have evolved to provide a robust understanding of risk; and our evolution to a patient-centric model demands that we leverage these methods more broadly.

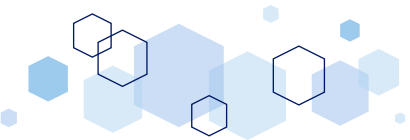
Lim et al. (2018)

- It may be unethical to adminster placebo or we want to minimise exposure to placebo, especially in lifethreatening disease areas
- Hard to recruit diseases (such as rare disease) or sub-groups (for example pediatrics)
- For bridging/extrapolating/partially extrapolating where high quality data exists, and there is a good scientific understanding of the biology
- Even in larger disease areas where natural disease progression is well understood, we could potentially get better treatments faster to patients
- Regulators are open to exploring alternative study designs

Minimizing Patient Burden Through the Use of Historical Subject-Level Data in Innovative Confirmatory Clinical Trials: Review of Methods and Opportunities | Therapeutic Innovation & Regulatory Science (springer.com); *Complex Innovative Trial Design Meeting Program | FDA*

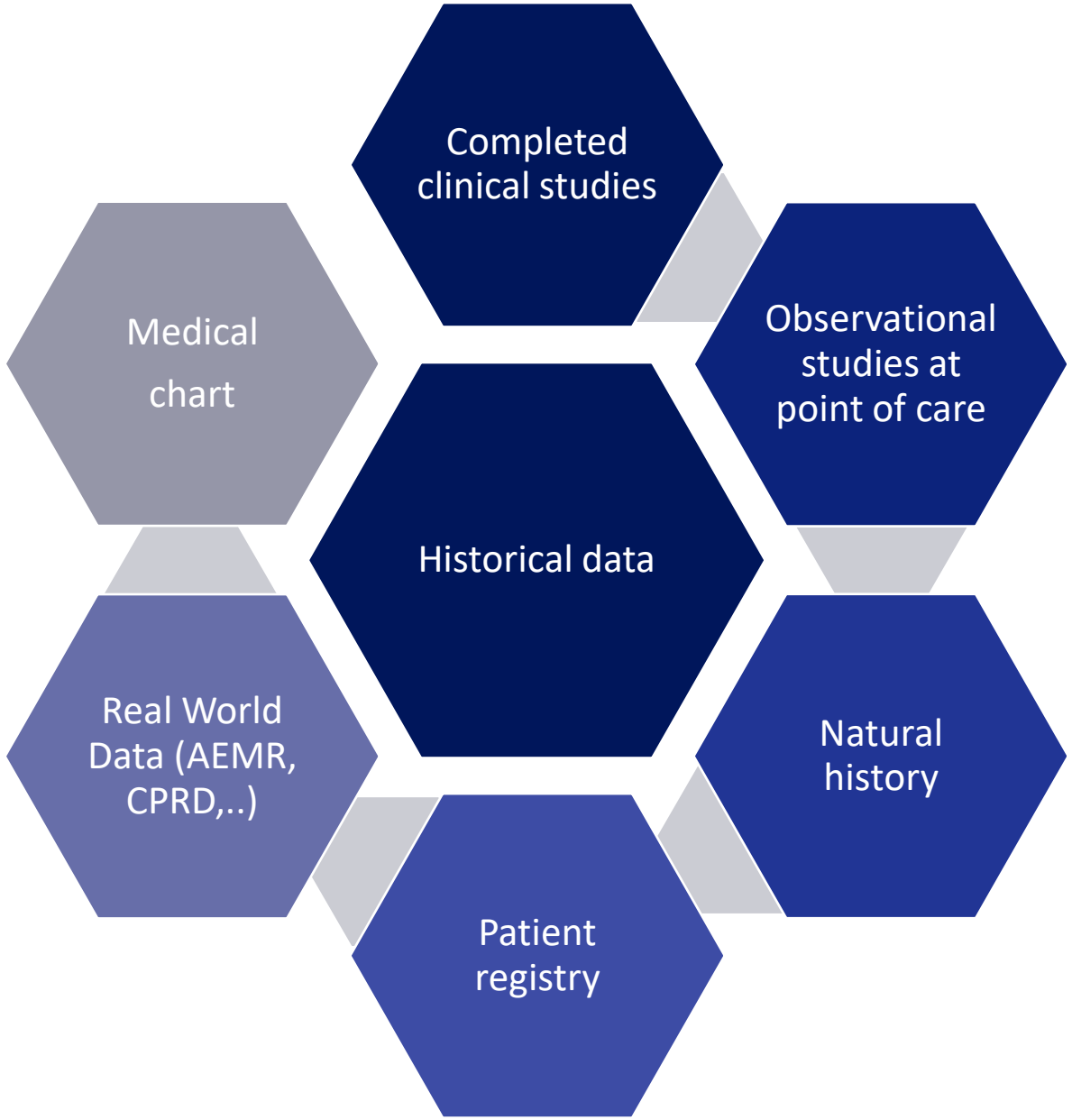
What’s not to like?

- Augmenting a RCT with historical data risks inflating the type I error in case of data conflict/drift/incongrunence between the concurrent control and the historical control
- This has to be weighed against the gain in power/precision



What to borrow?

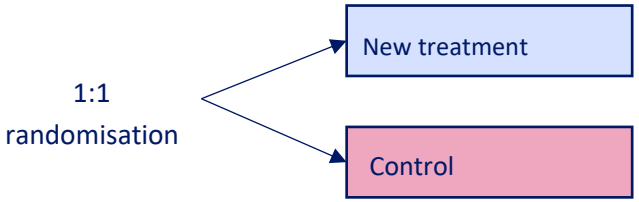
- In-house trial data (RCT, NIS)
- Systematic Literature Reviews
- **Data-sharing initiatives** (TransCelerate, Project Datasphere, Vivli, …)
- Several **cross-industry** working groups/consortia
  - Ex. European EFSPI/PSI Historical Data Special Interest Group, DIA Bayesian Working Group, Medical Device Innovation Consortium External Evidence Methods,…



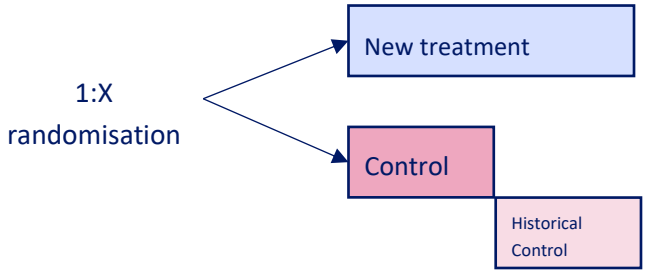
Bayesian Dynamic Borrowing in a nutshell

Bayesian Dynamic Borrowing in a ”nutshell”

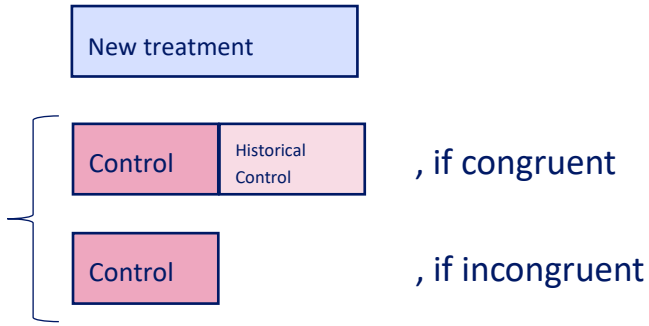
Standard RCT Design



BDB Design



Analysis (ideal)



What is the problem with ”borrowing”?

Scenario	New trt	Ctrl	Hist ctrl	Frequentist Operating Characteristics
Congruent	2	1	1	Increased power, increased type I error control
Incongruent, optimistic	1	1	0.5	Increased type I error
Incongruent, pessimistic	2	1	1.5	Reduced power

- Why might historical control data not align with concurrent data?
  - Population
  - Standard of care (sites, state of medical knowledge/reimbursement landscape)
  - Estimand
  - …

1. Such a group must have received a precisely defined standard treatment which must be the same as the treatment for the randomized controls.
2. The group must have been part of a recent clinical study which contained the same requirements for patient eligibility.
3. The methods of treatment evaluation must be the same.
4. The distributions of important patient characteristics in the group should be comparable with those in the new trial.
5. The previous study must have been performed in the same organization with largely the same clinical investigators.
6. There must be no other indications leading one to expect differing results between the randomized and historical controls. For instance, more rapid accrual on the new study might lead one to suspect less enthusiastic participation of investigators in the previous study so that the process of patient selection may have been different.

Only if all these conditions are met can one safely use the historical controls as part of a randomized trial. Otherwise, the risk of a substantial bias occurring in treatment comparisons cannot be ignored. For instance, 'literature' controls

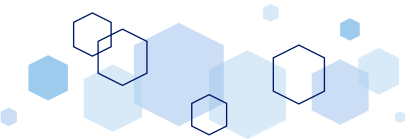
https://doi.org/10.1016/0021-9681(76)90044-8.

Elements of a BDB design

Elements of a BDB design: Bayesian Approach to Borrowing

- The Bayesian Analysis is an application of Bayes Theorem
- Suppose θ is a vector of parameter(s) of interest (for instance CFB in HbA1c)
- Let Y be a collection of ”data”
- $L(\theta|Y)$ is the likelihood
- $p(\theta)$ is our beliefs about the prior distribution of θ before seeing the data
- Our posterior belief after having seen the data

$$p(\theta|Y) = \frac{L(\theta|Y)p(\theta)}{\int L(\theta|Y)p(\theta)d\theta} \propto L(\theta|Y)p(\theta)$$



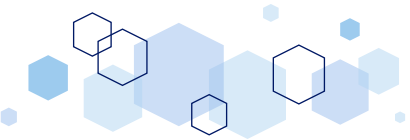
Elements of a BDB design: Incorporating historical data

- Suppose $p(\theta)$ is not assigned a vague prior, rather $p(\theta)$ is itself derived from historical data analysed with a vague prior say
- We have historical data D_0 and a prior $p_0(\theta)$

$$p(\theta|Y_0) \propto L(\theta|Y_0)p_0(\theta)$$

- Using Bayes rule again

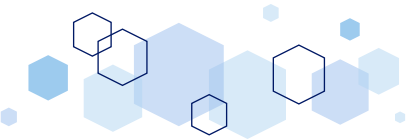
$$p(\theta|Y, Y_0) \propto L(\theta|Y)L(\theta|Y_0)p_0(\theta)$$



Elements of a BDB design: Bayesian Decision Rules

- Bayesian decision rules in clinical trials are based on the posterior density of the parameters of interest, for example:
  - **Superiority:** Declare trial a success if with 95% probability our posterior belief about θ exceeds some threshold value θ_0

$$\Pr(\theta \geq \theta_0 | Y) \geq 0.95$$



Elements of BDB Design:
Design Evaluation: Operating Characteristics (OC)

- Bayesian inference is based on the parameter space, as opposed to the frequentist approach (repeated sampling of the data)
- It is expected that the frequentist operating characteristics of the design is evaluated and presented

> "Pure" Bayesian approaches to statistics do not necessarily place the same emphasis on the notion of control of type I error as traditional frequentist approaches. There have, however, been some proposals in the literature that Bayesian methods should be "calibrated" to have good frequentist properties (e.g. Rubin, 1984; Box, 1980). In this spirit, as well as in adherence to regulatory practice, FDA recommends you provide the type I and II error rates of your proposed Bayesian analysis plan (see **Technical Details, Section 7**).

- Simulation based exploration of OC

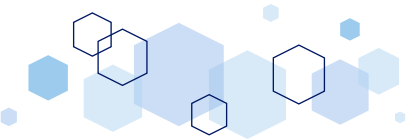
- Conditional or unconditional approach?
  - Should the OC be examined given the historical data at hand?
  - Or should we allow for both trial data and historical data to be sampled jointly?
- What is the sampling space for the historical data?
  - Excluding some historical data implies that we are potentially limiting the sampling space
- Distinction between *analysis prior* which is used when performing the final analysis, and ...
- the *design prior* which may explore an alternative set of assumption about the data, which is useful when evaluating the OC of the design

I (fda.gov)

Elements of a BDB design:
How much information is there in a prior? Effective sample size

- In order to be able to judge how many subjects ”worth” of information is embedded in the prior, the *prior effective sample size* can be useful
- Different proposals in the literature for a given prior $p(\theta)$
  - What is the sample size n that when combined with a minimally informative prior minimises the distance between the posterior and $p(\theta)$

Determining the Effective Sample Size of a Parametric Prior | Biometrics | Oxford Academic (oup.com)



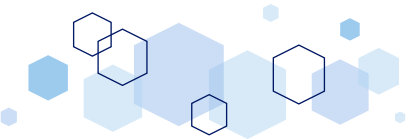
Proposals to incorporate historical data

- There are many proposals in the literature for how to downweigh historical evidence such that it does not override concurrent data
- Examples include:
  - Conditional power prior (static):

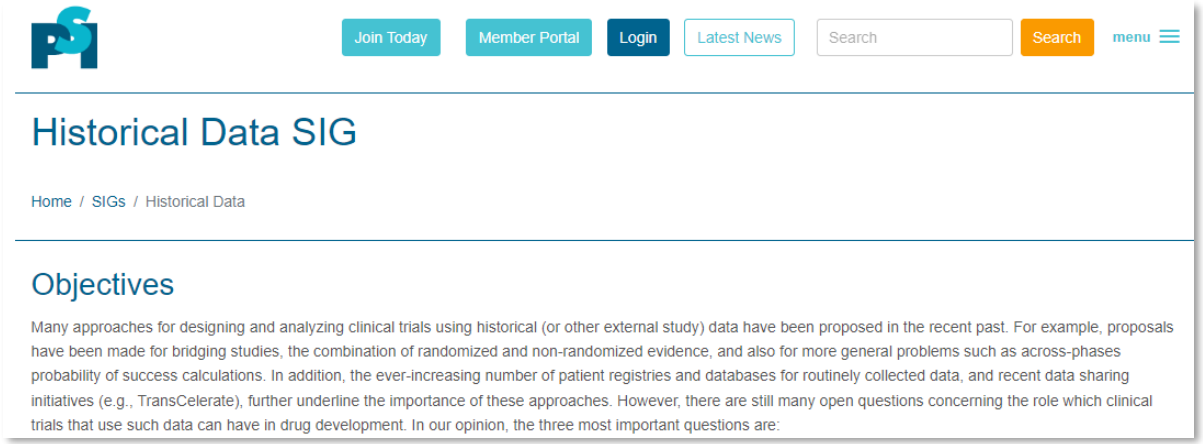
$$p(\theta|Y_0,\lambda) = \frac{L(\theta|Y)^{\lambda}p(\theta)}{\int L(\theta|Y)^{\lambda}p(\theta)d\theta}$$

    - Here $0 \leq \lambda \leq 1$ is chosen by the analyst. $\lambda = 0$ no weight on prior data; $\lambda = 1$ full weight
  - Elastic Prior: recent proposal to rapidly down-weigh historical data based on a congruence metric
  - Commensurate Prior: between study variation (dynamically) controls amount of borrowing
  - Robust MAP prior: Mixture prior of the meta-analytical predictive prior and a vague component

The power prior: theory and applications - Ibrahim - 2015 - Statistics in Medicine - Wiley Online Library ; *https://doi.org/10.1111/biom.13551* ; *https://doi.org/10.1214%2F12-BA722*; *https://onlinelibrary.wiley.com/doi/full/10.1111/biom.12242*



The robust MAP prior approach



- Introduced in Schmidli et al. (2014)
- The prior consists of two elements:
  - MAP = meta-analytical predictive prior
  - Robustification with a vague prior (= 1 subject), to limit type I error inflation in case of incongruence/drift

Historical Data (psiweb.org)
academic.oup.com/biometrics/article-abstract/70/4/1023/7419945

Full article: Beyond the Classical Type I Error: Bayesian Metrics for Bayesian Designs Using Informative Priors (tandfonline.com)

Constructing the MAP prior

- We have h = 1,...,H relevant historical studies of the control arm
- We synthesise the data in a random effects meta-analysis

$$Y_h|\psi_h \sim F(\psi_h)$$
$$\psi_h|\eta \sim G(\eta)$$
$$\eta \sim P$$

- F is the sampling distribution, G is the exchangeability distribution, P is a hyper-prior
- ψ_h are parameters (e.g. means of historical studies)
- η are parameters (e.g. between-study varation)

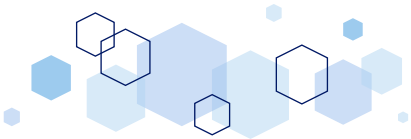
- Based on this synthesis we ask what is the predictive distribution for a new study (assuming exchangeability between historical data and a new study):

$$Y_c \sim F(\psi_c)$$
$$\psi_c|\eta \sim G(\eta)$$

- The MAP prior is defined as the marginal posterior distribution for ψ_c:

$$p_{MAP}(\psi_c) = p(\psi_c|Y_1, \ldots, Y_H).$$

- Two sources of variation: due to sampling and between-study variation



Robustification of the MAP prior

- To protect against incongruence/conflict we construct a mixture prior of the MAP and a vague prior

$$p_{RMAP}(\psi_c) = (1-w) \cdot p_{MAP}(\psi_c) + w \cdot p_V(\psi_c)$$

- The weight $0 < \omega < 1$ can be interpreted as the belief that the historical data is not relevant

- When the final analysis is conducted the congruence between the historical information and the concurrent control dynamically leads to updating of the weight:

$$p_{RMAP}(\psi_c|y_c) = (1-\tilde{w}(y_c)) \cdot p_{MAP}(\psi_c|y_c) + \tilde{w}(y_c) \cdot p_V(\psi_c|y_c)$$

where

$$\begin{aligned}(1-\tilde{w}(y_c)) &= \frac{(1-w) \cdot g_{MAP}(y_c)}{(1-w) \cdot g_{MAP}(y_c) + w \cdot g_V(y_c)}, \\ \tilde{w}(y_c) &= \frac{w \cdot g_V(y_c)}{(1-w) \cdot g_{MAP}(y_c) + w \cdot g_V(y_c)}\end{aligned}$$

and g denotes the marginal likelihood functions of the new data under either the MAP-prior or the vague prior:

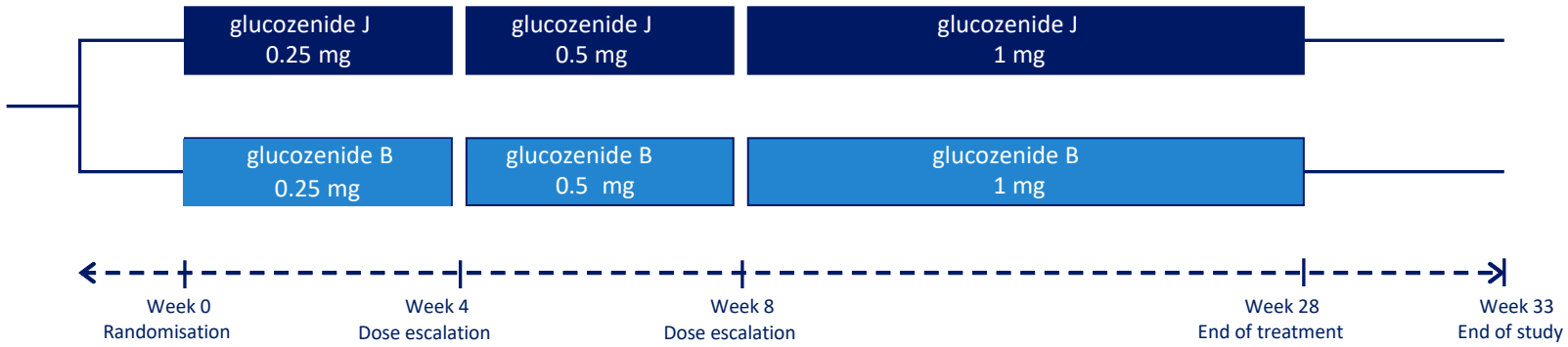
$$\begin{aligned}g_{MAP}(y_c) &= \int_\Psi f(y_c|\psi_c) \cdot p_{MAP}(\psi_c)\, d\psi_c, \\ g_V(y_c) &= \int_\Psi f(y_c|\psi_c) \cdot p_V(\psi_c)\, d\psi_c.\end{aligned}$$

A case from Novo Nordisk

Study Proposal (anonymized)

332 participants randomised 3:1

- T2D
- >=18 years old
- Metformin
- HbA1c 7-10.5%
- BMI (≥ 25 - ≤35 kg/m^2)
- eGFR ≥ 60 mL/min/1.73 m2

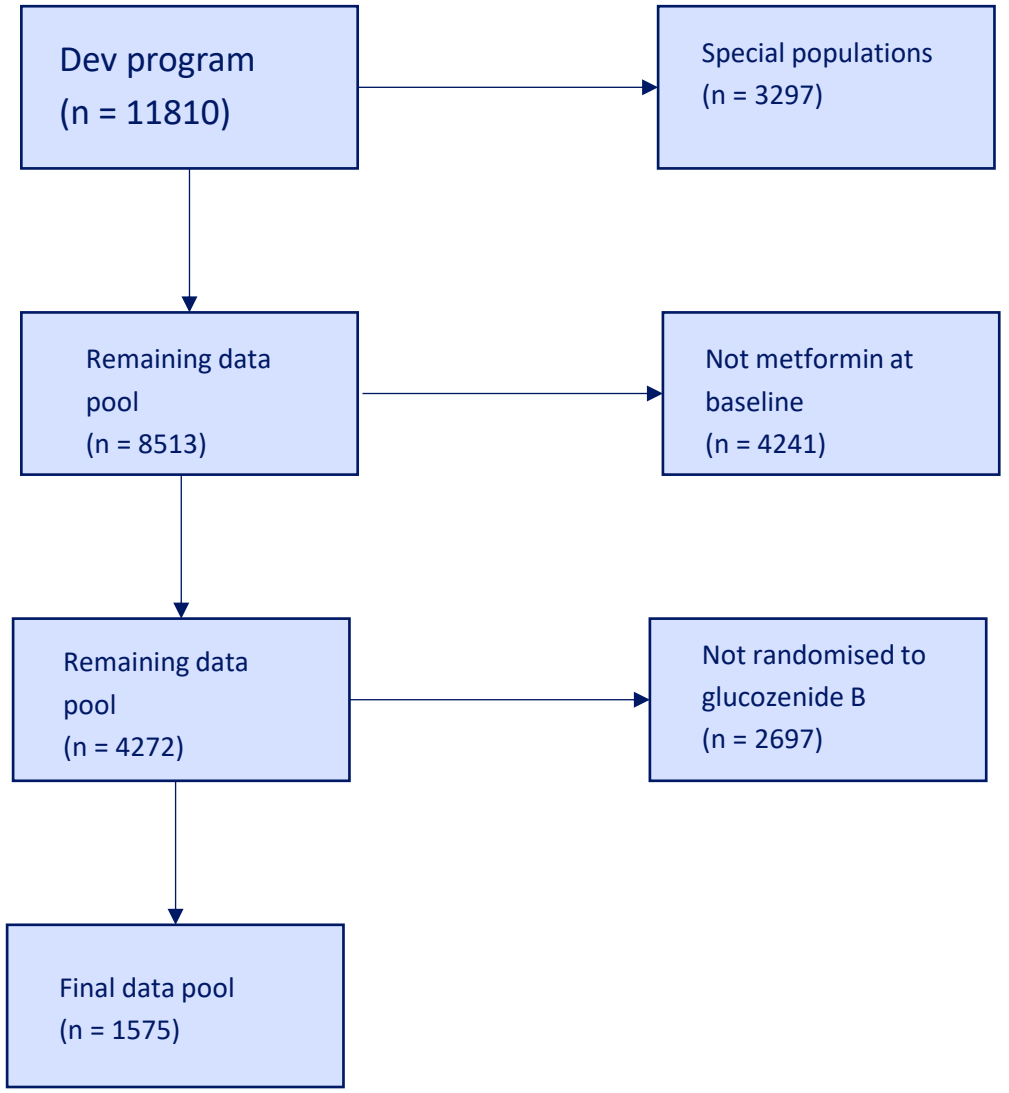


Trial objective	Key endpoints	Other information

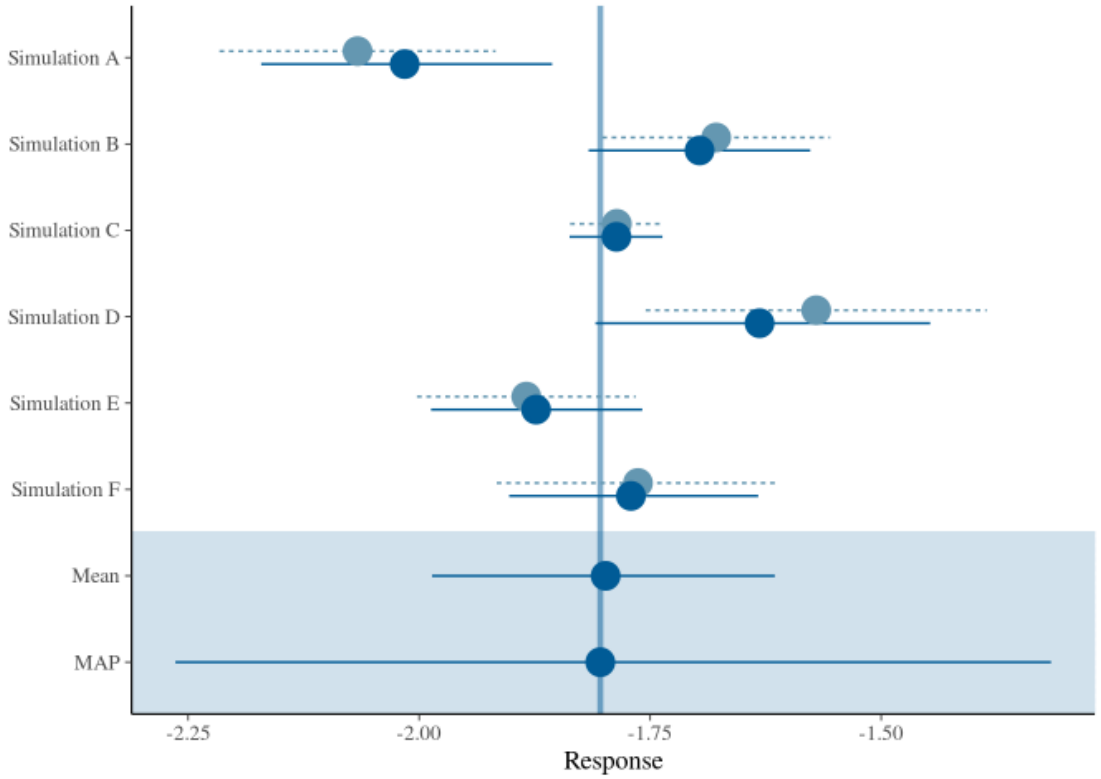
Demonstrate clinical comparability between glucozenide J and glucozenide B	• **Primary:**
	  • Change from baseline in HbA_{1c}
	• **Secondary:**
	  • Body weight
	  • Adverse Events
	  • Anti-drug antibody (%)
		• Double-blind, active-controlled
		• Comparability margin: 0.3%-points
		• Sparse PK sampling
		• Primary assessment is augmented with historical data

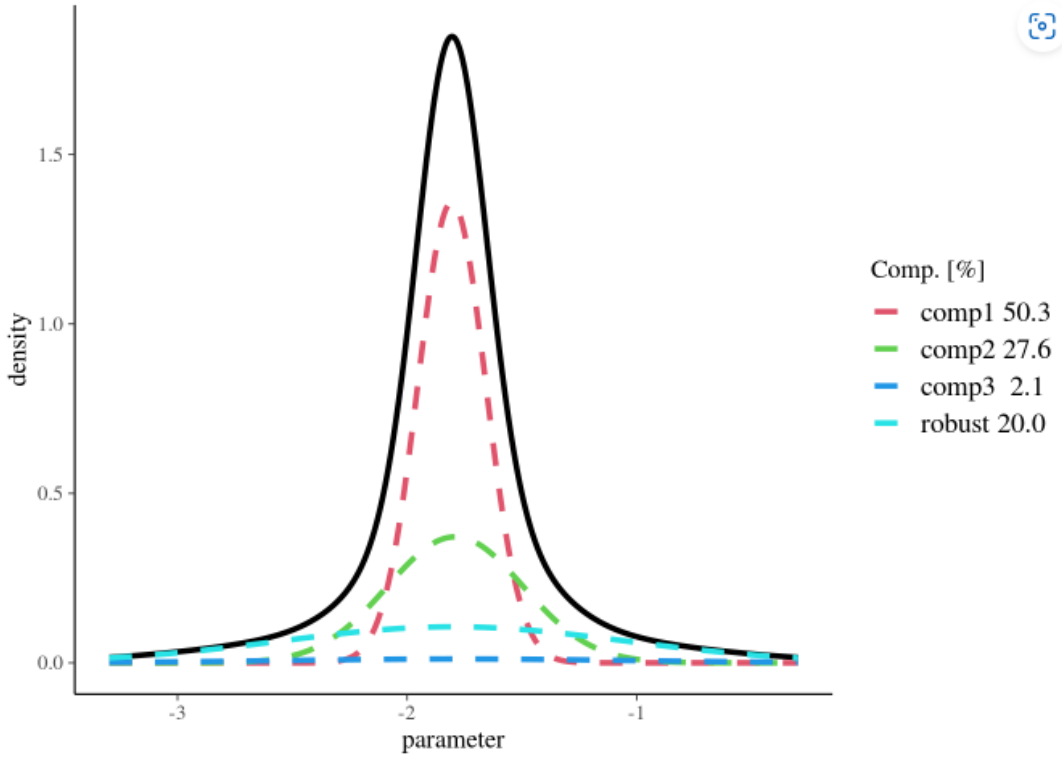
Participant selection and Summarising Historical Evidence

- Development Programme
  - glucozenide phase 3
  - Same treatment (glucozenide B)
  - Same sponsor
- HbA_{1c} is an objective outcome measure
- HbA_{1c} evaluated at week 28 or 30



The MAP and the robustified MAP prior





Synthetic data closely aligned to actual study data

Light blue – dashed: study results with 95% CI. Dark blue – solid: Shrunk estimates with 95% CrI; Mean and MAP with 95% CrI

Proposed design

- The MAP prior represents information from about 20 subjects (with an assumption of population SD being the same in historical and new trial)
- After robustification this is reduced to around 14 subjects
- 3:1 randomisation, 249 subjects to investigational treatment and 83 to comparator treatment
- This suggest that prior information will not dominate the new trial

- Proposed decision rule:

To evaluate the operating characteristics we need to set a decision rule and an expectation for the investigational treatment in the new study. In this example we wish to confirm that the investigational treatment is non-inferior to the comparator on HbA1c and so we set the criterion to $P(\theta_{act} - \theta_{comp} < 0.3) > 0.975$.

- Explore scenarios where data conflict is ± 0.5

Operating Characteristics (for the chosen design)

Maximum
Type I error

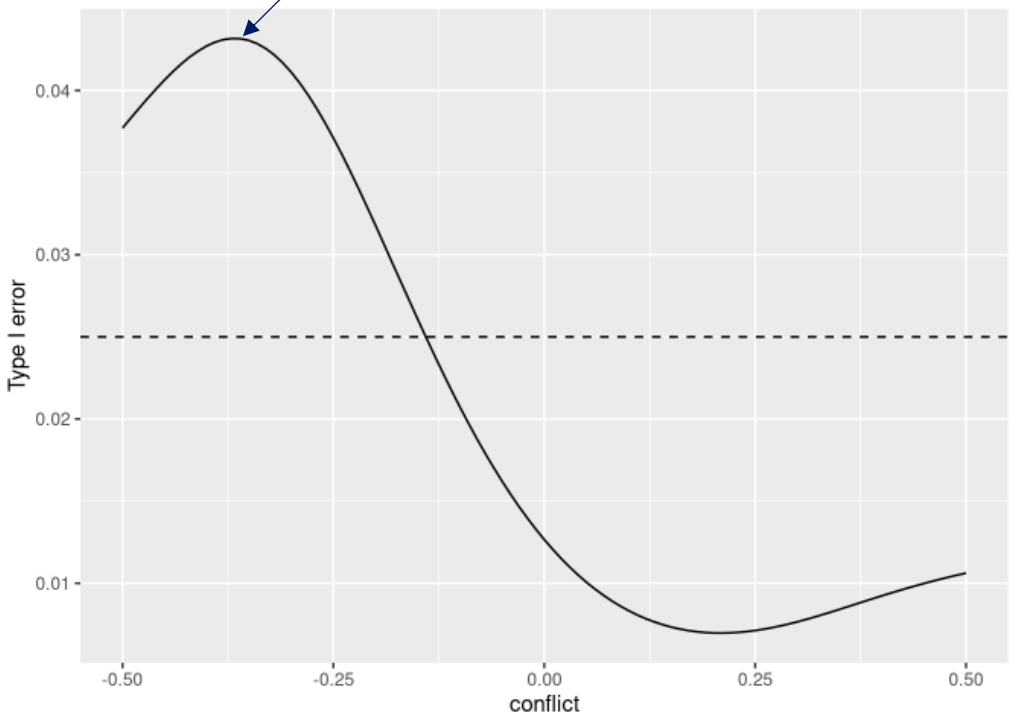


Figure 13.4: Type I error sensitivity to prior-data conflicts. Dashed line represents the target type I error rate.

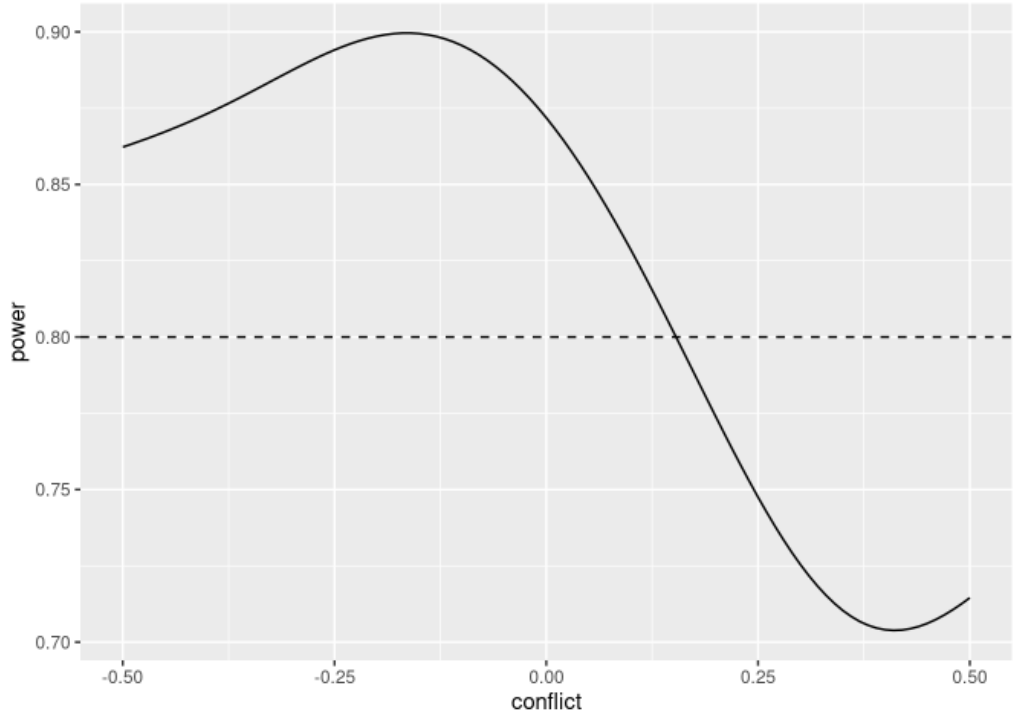


Figure 13.5: Power sensitivity to prior-data conflicts. Dashed line represents the target power.

Sweetspot

Sweetspot: range of conflict where type I error is controlled and where power is above target power

Regulatory Feedback & Learnings

- In this particular case the FDA was not accepting of the study proposal to augment with historical data for the primary analysis
- Other agencies were more open to the suggested approach

- Essentially all of Pocock's criteria were satisfied
- The design had inflated type I error in some areas of the sampling space (type I error not uniformly controlled)
- The disease is not rare
- Not sufficient rationale for alternative bias/variance trade-off

Regulatory Acceptance

Ultimately a review issue ...

In general, the acceptance of Bayesian methods in clinical trial research goes slow. The regulators worry that the prior distributions are based on favorable data for the experimental drug, so that bias and an inflated Type I error rate are caused. As a result, the Bayesian approach is in general tolerated and accepted only when conventional clinical trial designs are impossible to implement in practice. This is the case for orphan diseases and pediatric studies, as mentioned in the Introduction. For medical devices the Bayesian approach is generally accepted by the regulatory authorities assuming, of course, the same rigorous setup and conduct as with a frequentist approach, see, for example, Haddad (2020). Moreover, the Bayesian approach often is the recommended approach to deal with small sample issues.

- However regulators are increasingly willing to engage in discussion around more complex designs, including designs that borrow from historical data

https://doi.org/10.1214/24-BJPS598

CID Case Study: Pediatric Patients with MS

CID Case Study: A Study in Pediatric Patients with Multiple Sclerosis

Study Design:

The proposed study is a randomized, double-blind, Bayesian, group sequential, non-inferiority (NI) trial comparing an investigational treatment to an active control in pediatric patients with multiple sclerosis (MS), borrowing strength from external data in adults and children. The primary endpoint is the annualized relapse rate (ARR). One interim analysis for efficacy is planned.

The available external studies consist of a completed trial in pediatric MS patients and several trials in adult MS patients. A Bayesian framework will be used to incorporate the information from these studies using informative meta-analytic predictive (MAP) priors for the parameters of the statistical model. This MAP prior is combined with a non-informative prior component to produce a robust meta-analytic predictive (RMAP) prior that adapts the amount of information being borrowed based on the compatibility between the prior and observed data.

Discussion:

A topic of the discussion involved the benefit of the current NI design over a superiority trial given the treatment landscape of pediatric multiple sclerosis, a rare disease with an unmet need. A non-inferiority design would not expose pediatric MS patients to placebo, given the existence of an approved treatment. The sponsor and Agency acknowledged that a NI trial may be more attractive to pediatricians and patients and may potentially minimize patient burden.

The discussion also centered on a feasible and appropriate margin. The Agency requested a comprehensive and systematic literature review to justify the non-inferiority margin taking between-trial heterogeneity into account. The Agency recommended that a cautious approach to NI margin selection was warranted given that only a single historical trial was conducted in the pediatric MS population and there was uncertainty about using adult study findings for extrapolation. The FDA also suggested exploration of a modeling strategy incorporating additional, relevant data and accounting for the differential treatment effect by age. Moreover, estimation of the effect of the active comparator should incorporate data from controlled studies and between-study variability should be modeled. The FDA indicated the importance of consistency in the effect between trials. Additionally, FDA stated that the sponsor should adequately address the statistical implications of using the same historical data to inform both the NI margin and the prior. The Agency requested extensive simulations regarding the proposed priors and operating characteristics of the planned design.

CID Case Study: A Study in Pediatric Patients with Multiple Sclerosis (fda.gov)

CID Case Study: External Control in oncology

CID Case Study: External Control in Diffuse B-Cell Lymphoma

Study Design:

The proposed trial is a randomized, open-label, multicenter trial in patients with first-line diffuse large B-cell lymphoma. Patients are to be randomized 2:1 to treatment vs. control. The primary endpoint of the study is Investigator-assessed progression-free survival (PFS), defined as the time from randomization to the first occurrence of progression or relapse, using the 2014 Lugano classification for Malignant Lymphoma (Cheson et al. 2014), or death from any cause, whichever occurs first.

The key secondary endpoint is overall survival (OS). The analysis population for OS will be augmented by patients from an external control arm so that approximately half of the patients in the resulting control group are comprised of patients from the external control. The external control arm will be partially concurrent with the planned trial. The planned analysis of OS utilizes a Bayesian commensurate prior with a Weibull model (Lewis et al. 2019) to dynamically borrow information from the external control arm. Furthermore, propensity score matching will be conducted to select external control patients for inclusion in the analysis. Inference will be based on the posterior mean and 95% credible interval of the posterior distribution of the hazard ratio.

Discussion:

Innovative designs, such as those proposed under the CID program, often require stronger assumptions than designs commonly considered for regulatory decision-making. In addition, key operating characteristics such as power and Type I error may not have closed-form analytical expressions. Consequently, simulations are necessary to understand the operating characteristics of these designs. In this case, the Sponsor provided simulations to understand these operating characteristics in the case of violations from the various model assumptions, namely the proposed Weibull distribution, the linear form of the propensity score model, the assumption of no unmeasured confounding, and the assumed similarity in patient populations. These simulations facilitated discussion between the Sponsor and FDA on modeling choices and practical considerations for assessing the results.

In general, FDA prefers trial designs and analyses which require minimal assumptions and which result in straightforward interpretation of the treatment effect in the associated population. In this case, a consideration was whether the propensity score could be used as a covariate in the Weibull model for overall survival. In this case, FDA believed that use of the propensity score as a covariate would make results difficult to interpret and communicate. Consequently, the Sponsor specified propensity score matching as the method for adjusting for baseline differences in populations.

While simulations are important for assessing operating characteristics intractable to analytical assessment, many model assumptions are more tenable if supported by expert clinical input, historical data, or thoughtful plans for trial implementation. For instance, the chosen covariates for the propensity score model required clinical rationale based on expert clinical opinion and literature. In addition, the assumption of the Weibull distribution was supported by results from trials in this disease area which appeared to be reasonably fit with the Weibull distribution. The assumption of similarity in patient populations was bolstered by the Sponsor's plan to prioritize enrolling patients in the same sites for both the randomized arms and external control arm when possible. While rationale and simulations provide crucial support in designs with strong assumptions, ultimately many of these assumptions are unverifiable. Careful review of the final results will be necessary to further understand the strengths and limitations of this design.

https://www.fda.gov/media/155405/download?attachment

FDA review paper on CID (Price and Scott, 2021)

- CIDs (including BDBs) may be considered when:
  - Clear unmet need
  - Conventional methods not feasible/optimal
  - Proposed Methods are reliable
    - Explore operating characteristics (via simulations); particularly for scenarios where there is drift

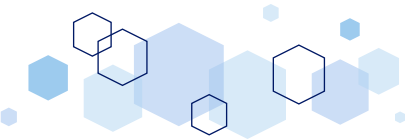
An overarching goal of using complex innovative trial designs is to improve clinical trial efficiency with scientific methods that can reliably answer the questions of interest and facilitate regulatory decisions. Clinical trial efficiency may translate into a reduction in numbers of patients needed for a trial, accelerated product development, or optimized product development (e.g. maximum information obtained from the research effort). Complex innovative trial designs may be especially promising when conventional approaches may not be feasible or optimal, such as in areas where the population size is small or limited or where there is an unmet medical need.[3] Specifically, for small populations, design innovations that can reduce sample size may not only speed development but also make infeasible development programs feasible. In the setting of an unmet medical need where a conventional trial may not be feasible or optimal, a complex innovative design may result in accelerated product development and earlier product availability to patients.

The U.S. Food and Drug Administration's Complex Innovative Trial Design Pilot Meeting Program: Progress to date (sagepub.com)

Summary

Summary

- Strong rationale for using historical controls, including:
  - Address underserved populations
  - Getting treatments faster to patients
  - Where it is unethical to give placebo, or want to minimise exposure to placebo
  - High unmet medical need
- Regulatory adoption of historical borrowing designs like BDB is slow, with good reason:
  - Type I error is not uniformly controlled
- Maximise acceptance:
  - Clear rationale for why borrowing is needed
  - Engage early with authorities
  - Transparent selection of historical data and transparency of assumptions
  - Explore operating characteristics of the design in plausible parts of the sampling space
  - Clear reporting to allow assessment of the influence of historical data for the result
- Part of the difficulty may be in communicating consistently with regulators, to allow assessment of the risk/benefit of the proposed design
- PSI Historical Data SIG seeking qualification opinion to EMA on a framework for BDB
- FDA has a commitment to publish draft Guidance on the Use of Bayesian Methodology in Clinical Trials of Drugs and Biologics by September 30, 2025



References

References

- A review of dynamic borrowing methods with applications in pharmaceutical research (projecteuclid.org)
- https://onlinelibrary.wiley.com/doi/full/10.1111/biom.12242
- https://doi.org/10.1002/sim.9095
- Minimizing Patient Burden Through the Use of Historical Subject-Level Data in Innovative Confirmatory Clinical Trials: Review of Methods and Opportunities | Therapeutic Innovation & Regulatory Science (springer.com)
- Elastic priors to dynamically borrow information from historical data in clinical trials - Jiang - 2023 - Biometrics - Wiley Online Library
- Complex Innovative Trial Design Meeting Program | FDA
- The U.S. Food and Drug Administration's Complex Innovative Trial Design Pilot Meeting Program: Progress to date - Dionne Price, John Scott, 2021 (sagepub.com)
- https://doi.org/10.1002/jcph.2132
- Full article: Beyond the Classical Type I Error: Bayesian Metrics for Bayesian Designs Using Informative Priors (tandfonline.com)
- The power prior: theory and applications - Ibrahim - 2015 - Statistics in Medicine - Wiley Online Library
- https://doi.org/10.1214%2F12-BA722
- https://doi.org/10.1016/0021-9681(76)90044-8
- Determining the Effective Sample Size of a Parametric Prior | Biometrics | Oxford Academic (oup.com)

Advancing Precision Medicine with Innovative In Silico Approaches in Reproductive Medicine

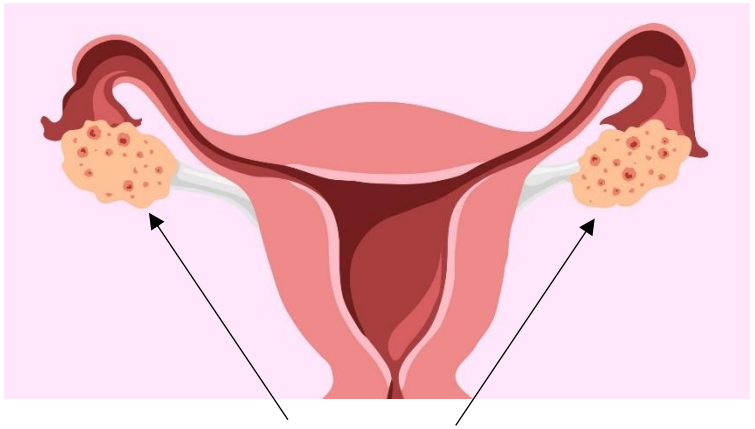
Daniël Jonker

Director of Clinical Pharmacology

Early Sciences

FERRING PHARMACEUTICALS

Introducing a few terms from reproductive medicine...



Ovaries
(containing follicles)

Oocyte

In vitro fertilization
(IVF)

Follicle stimulating
hormone (FSH)

Can ovarian physiology be accurately modelled to generate novel insights and guide the optimal use of FSH?

FERRING
PHARMACEUTICALS

Evolution of modelling approaches to guide the optimal use of follicle stimulating hormone in *in vitro* fertilisation

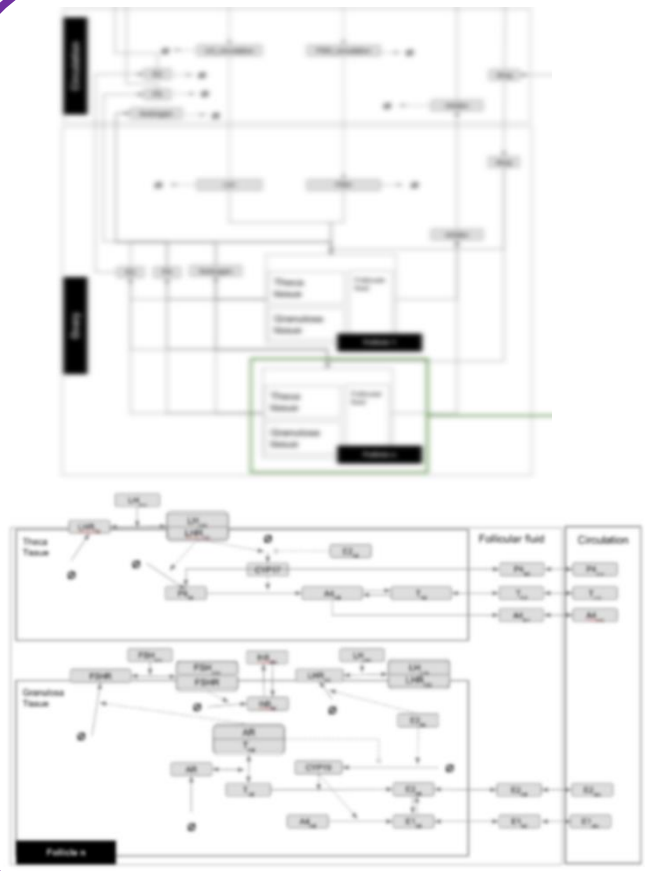
Descriptive model for oocytes retrieved

$$\lambda = \frac{E_{max} \cdot Dose^{\gamma}}{ED_{50}{}^{\gamma} + Dose^{\gamma}}$$

Covariates:
E_{max} : AMH
Dose: body weight

Individualised dosing algorithm for better safety

Physiological model



Understand and predict how the ovary responds to drug interventions

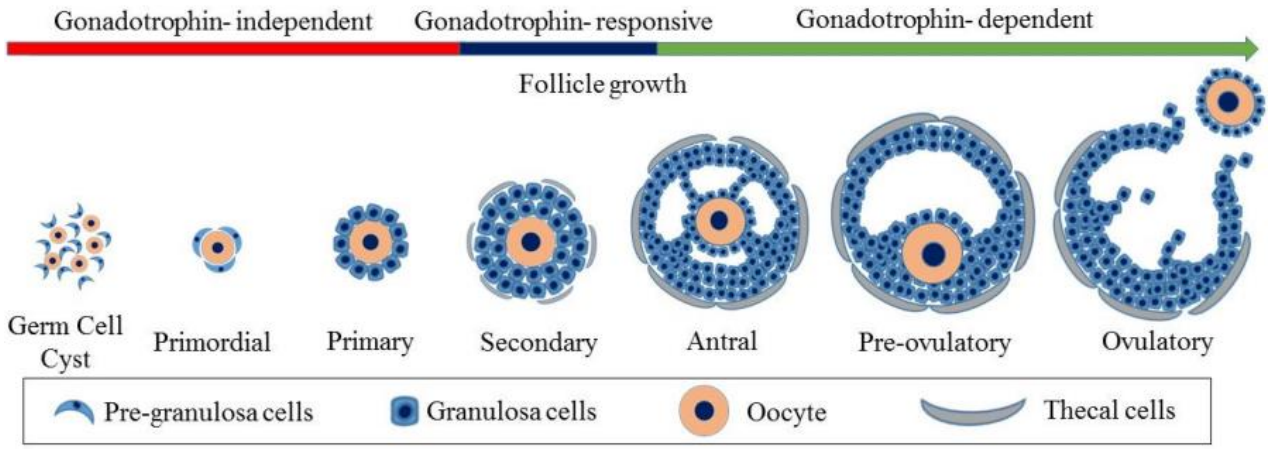
Inform dosing regimens and inclusion/exclusion criteria for clinical trials.

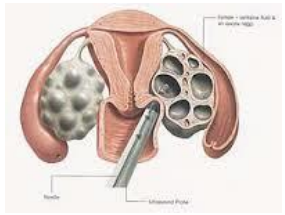
Potential application: individualised forecasts of IVF outcome.

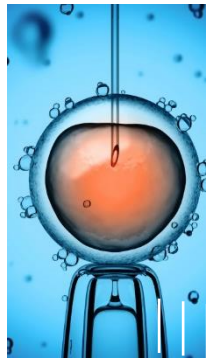
IVF: in vitro fertilisation
AMH: anti-müllerian hormone

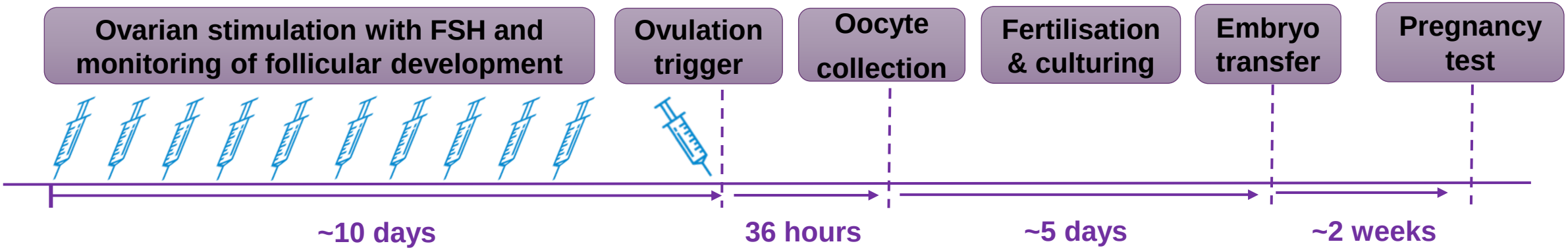
In vitro fertilisation (IVF) in a nutshell

Stimulating multifollicular growth with the aim to achieve pregnancy







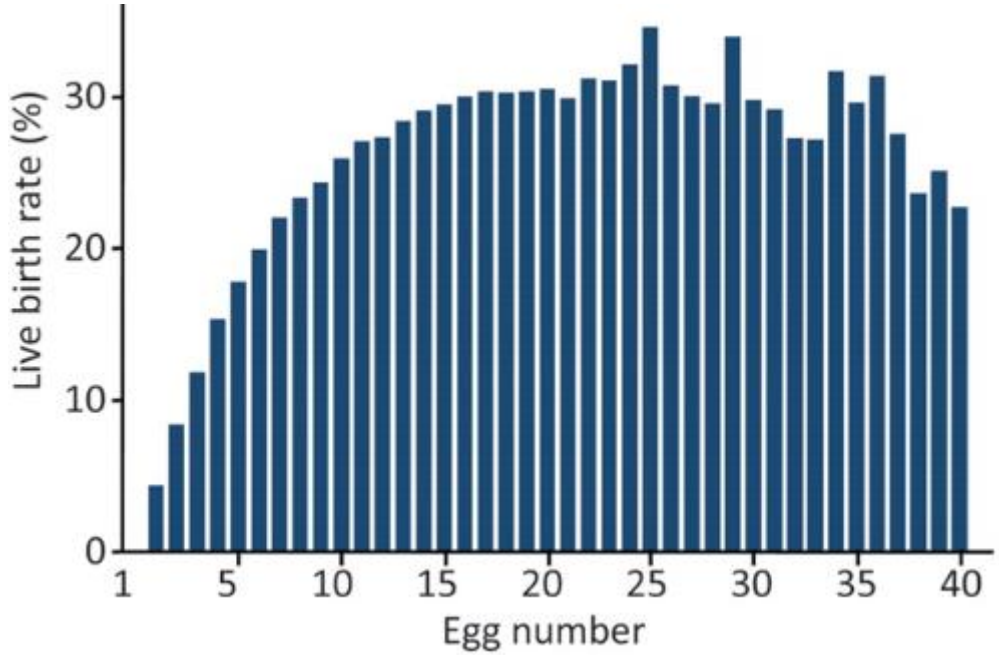


FSH: follicle stimulating hormone; one of the gonadotropins.

Sharum, Isam. (2016). Regulation of TGFβ/Smad Signalling During Early Follicle Development in the Mouse Ovary. 10.13140/RG.2.2.24992.02567.

FERRING PHARMACEUTICALS

Number of oocytes can be highly variable



Anti-müllerian hormone (AMH) is a key predictor for ovarian response

Low AMH

High AMH

Importance of control of stimulation

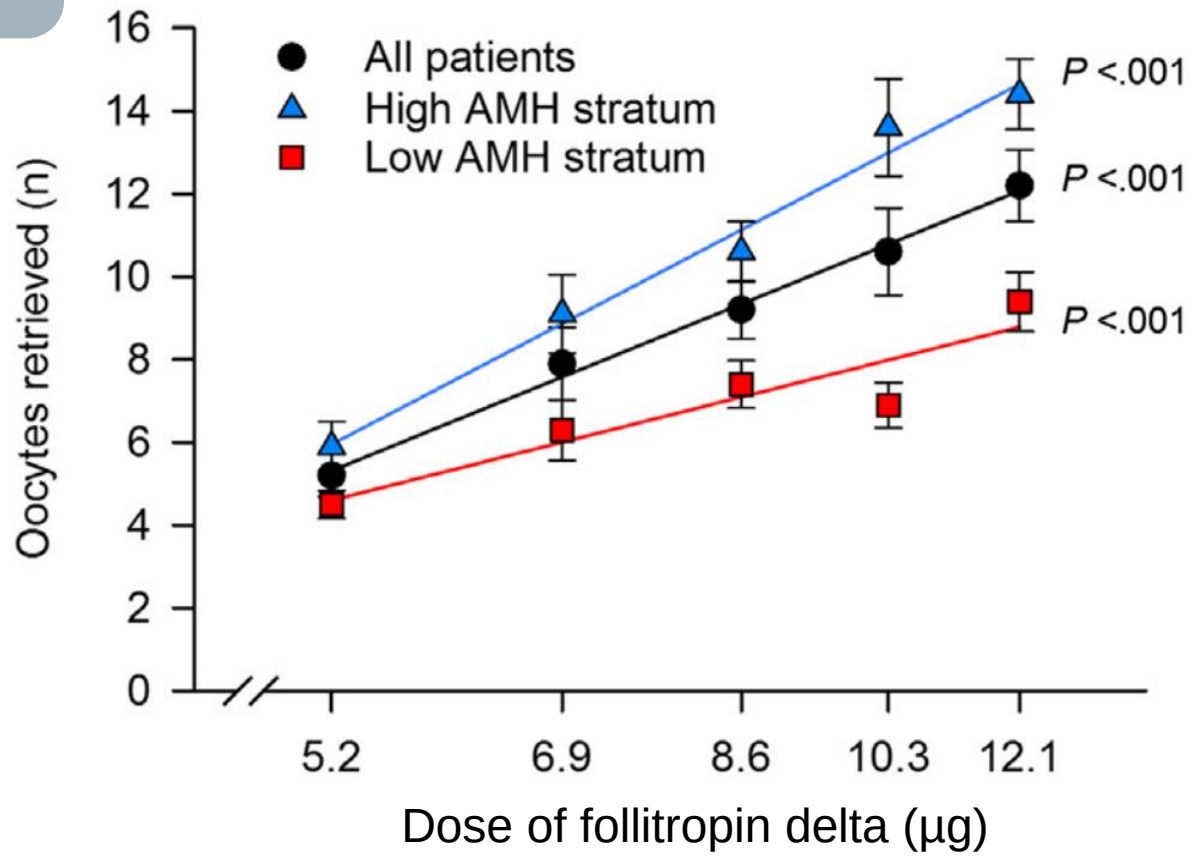
- Balance between too few oocytes and risk of ovarian hyperstimulation syndrome (OHSS)
- Invasive and costly procedure – need to create best chance for pregnancy at 1st attempt

Sunkara SK, et al. Hum Reprod. 2011 Jul;26(7):1768-74. doi: 10.1093/humrep/der106.

Follitropin delta dose finding trial in patients

Follitropin delta
- A recombinant FSH derived from a human cell line.
- Pharmacokinetics are different from other FSH.

- Women undergoing IVF were randomised to 1 of 5 dose levels of follitropin delta.
- The number of oocytes retrieved increased with the dose of follitropin delta.
- AMH level also significantly affected the number of oocytes retrieved.
- Patients with high AMH will require a lower dose of follitropin delta than patients with low AMH.



IVF: In vitro fertilisation
AMH: anti-müllerian hormone

Arce JC, et al. Fertil Steril. 2014 Dec;102(6):1633-40.e5. doi: 10.1016/j.fertnstert.2014.08.013

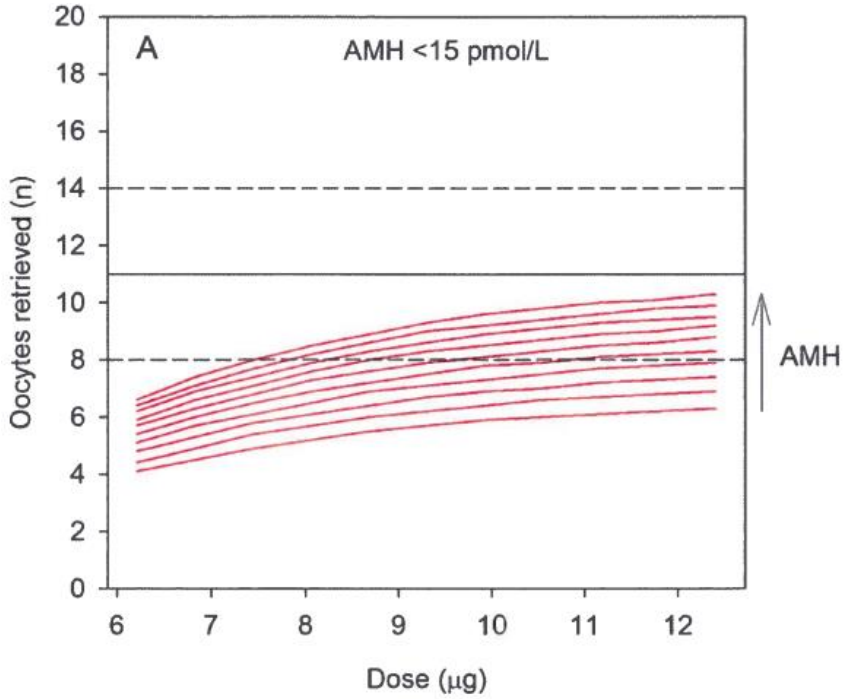
An individualised dosing algorithm was developed based on a descriptive dose-response model for oocytes retrieved

Dose-response model

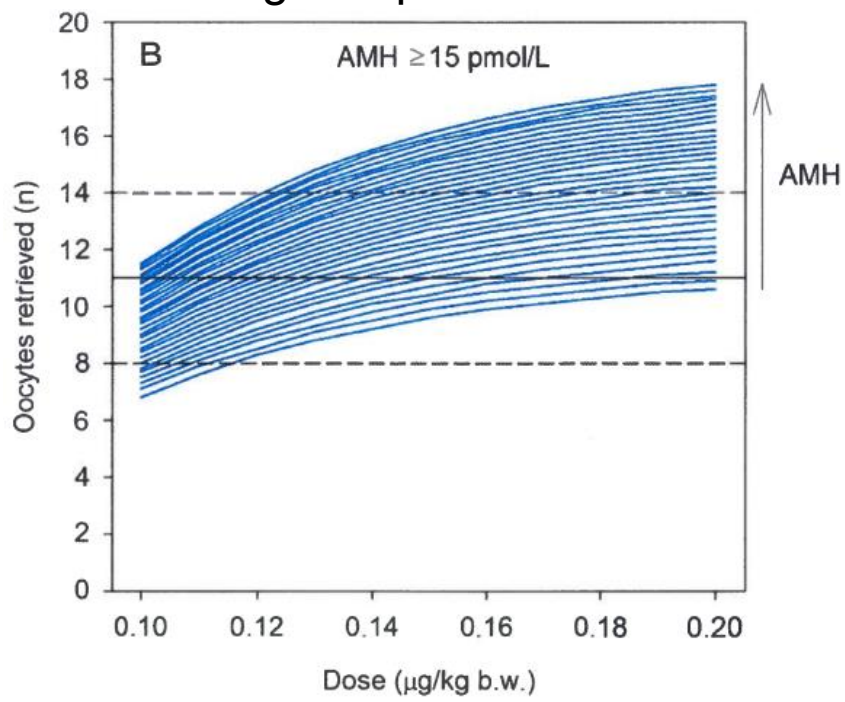
$$\lambda = \frac{E_{max} \cdot Dose^{\gamma}}{ED_{50}{}^{\gamma} + Dose^{\gamma}}$$

Covariates:
E_{max} : AMH
Dose: body weight

‘Low responders’



‘High responders’



Outcome

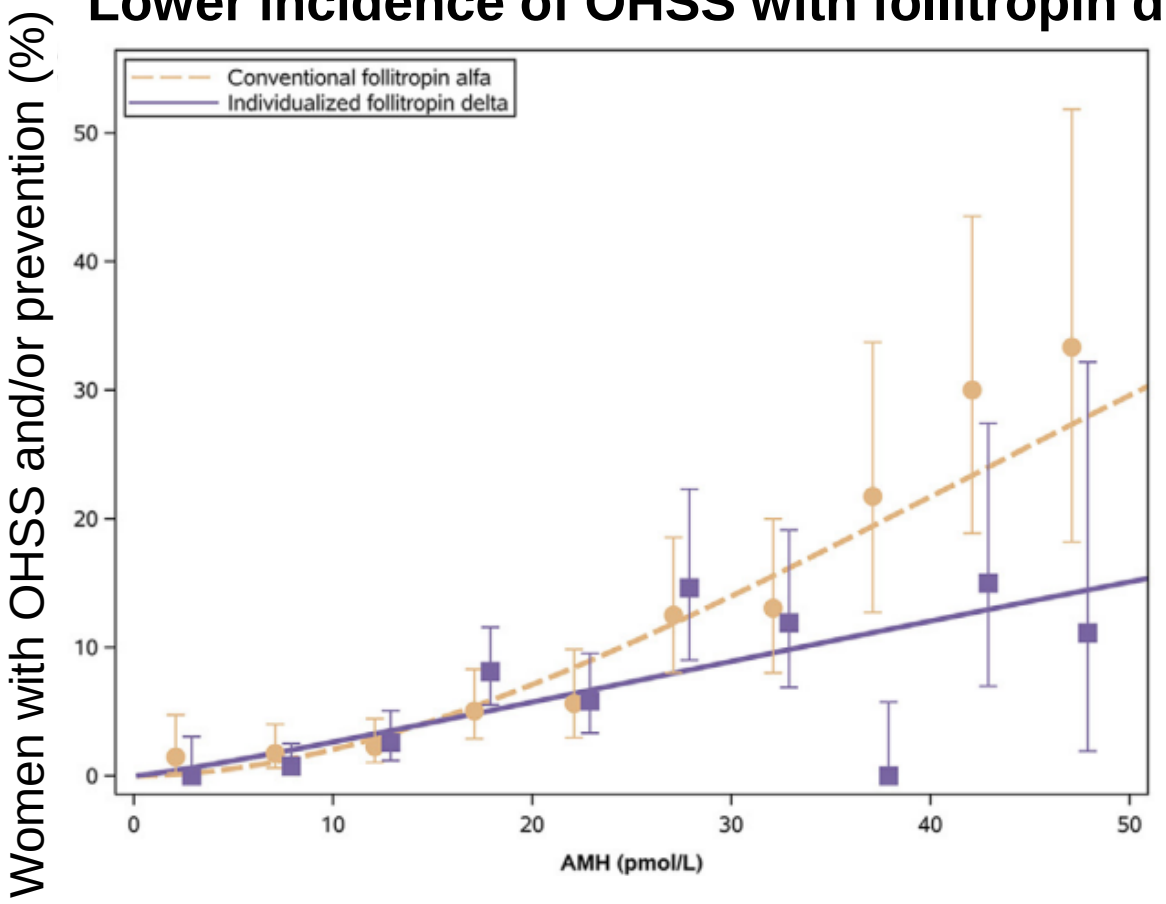
Follitropin delta dosing algorithm in 1st treatment cycle

AMH (pmol/L)	<15	15-16	17	18	19-20	21-22	23-24	25-27	28-32	33-39	≥40
Dose (µg/kg)	12 µg	0.19	0.18	0.17	0.16	0.15	0.14	0.13	0.12	0.11	0.10

Arce JC, et al. Using AMH for determining a stratified gonadotropin dosing regimen. In: Anti-Müllerian Hormone, 2016; Nova Science Publishers. Editors: Seifer DB and Tal R.

Individualised dosing afforded an improved safety profile with efficacy maintained in confirmatory clinical trials

Lower incidence of OHSS with follitropin delta



Nyboe Andersen A, et al. Fertil Steril. 2017 Feb;107(2):387-396.e4. doi: 10.1016/j.fertnstert.2016.10.033.

Non-inferior ongoing pregnancy rate

Follitropin delta (individualised)	Follitropin alfa (fixed starting dose)
30.7%	31.6%

Conclusion

- A simple mathematical equation made it possible to improve how FSH is dosed.

Limitations

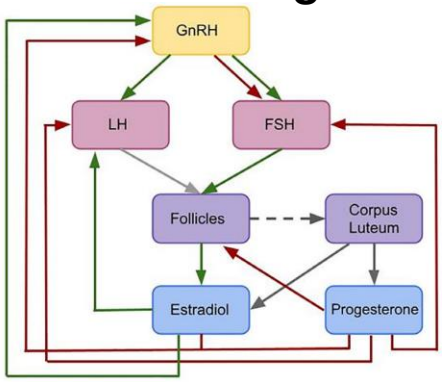
- The only inputs to the model are the daily dose, AMH level and body weight.
- The model does not enable predicting outcome of other patient characteristics, or effects of a different dosing frequency, or of dose changes during stimulation.

OHSS: ovarian hyperstimulation syndrome

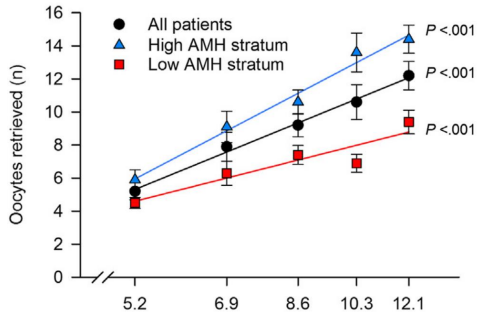
Building a new, more physiological model

The body of literature on follicle development is extensive

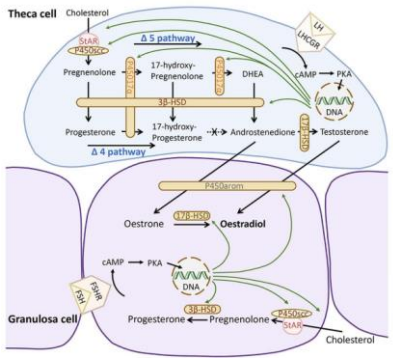
Hormonal regulation



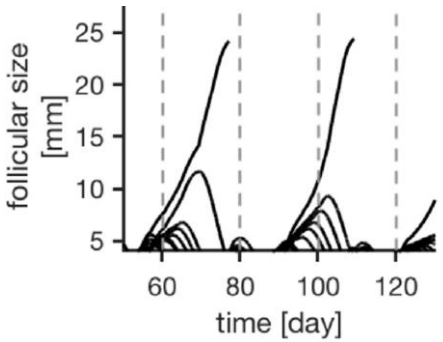
Clinical trial data



Cellular processes



Existing models



What to include?

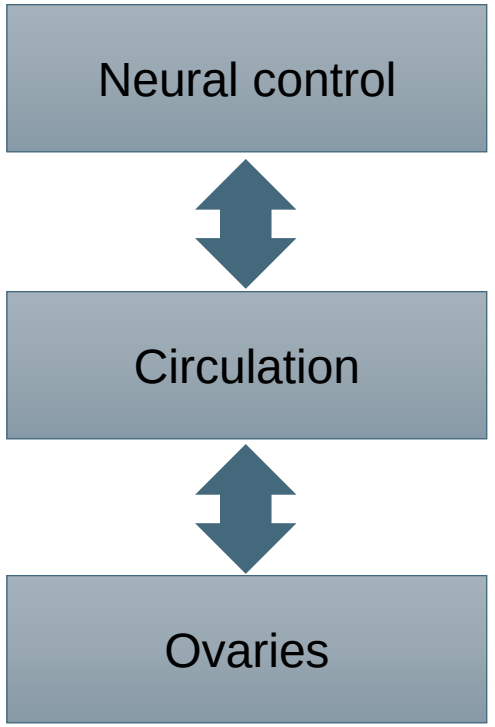
What level of detail?

Well-defined objectives

Which outcomes?

Which clinical settings?

Modular organisation



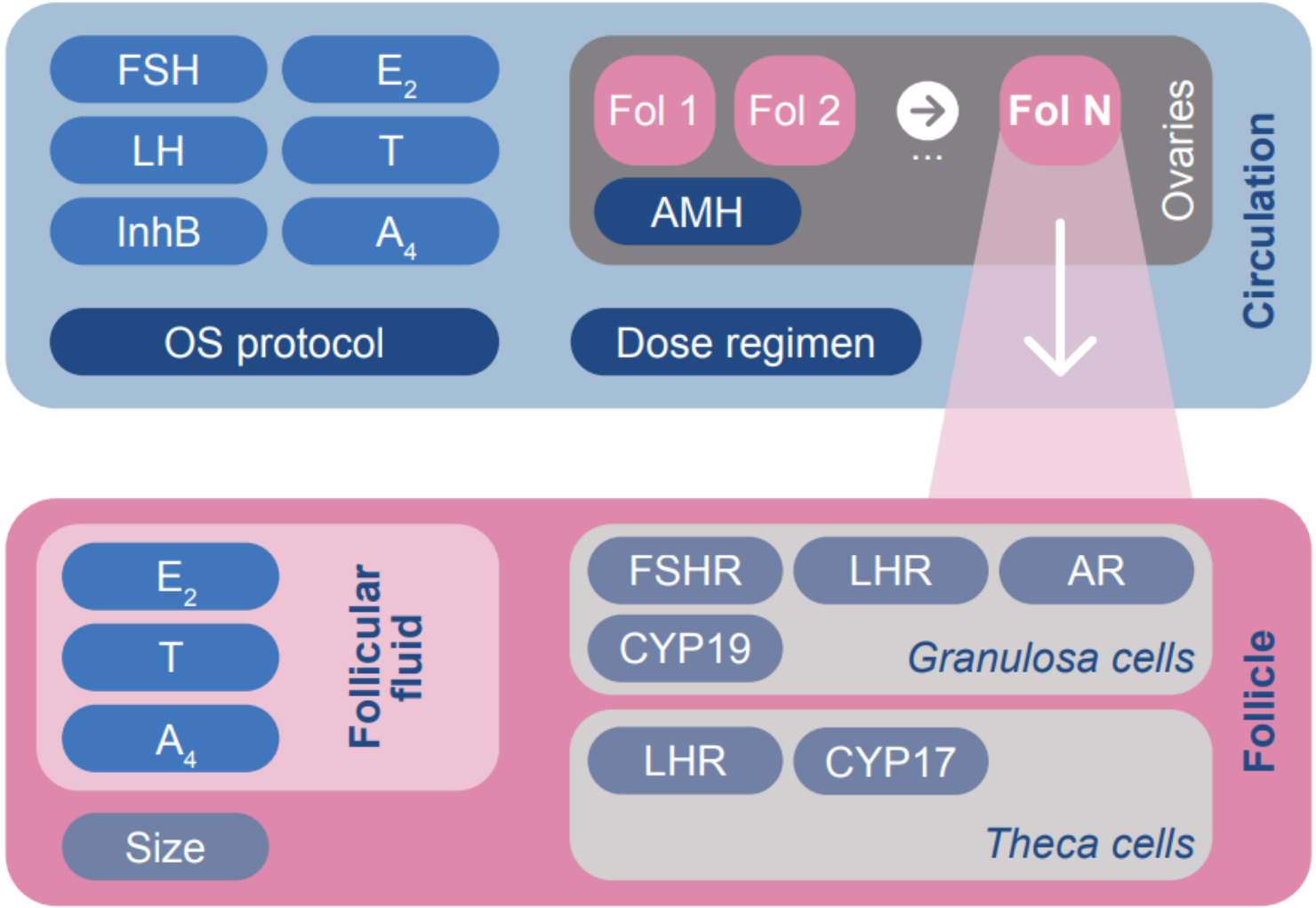
Zheng M, et al. Front Endocrinol. 14:1268248. doi: 10.3389/fendo.2023.1268248; Fischer S, et al. Front Endocrinol 12:613048. doi: 10.3389/fendo.2021.613048.

Computational model framework

Processes spanning

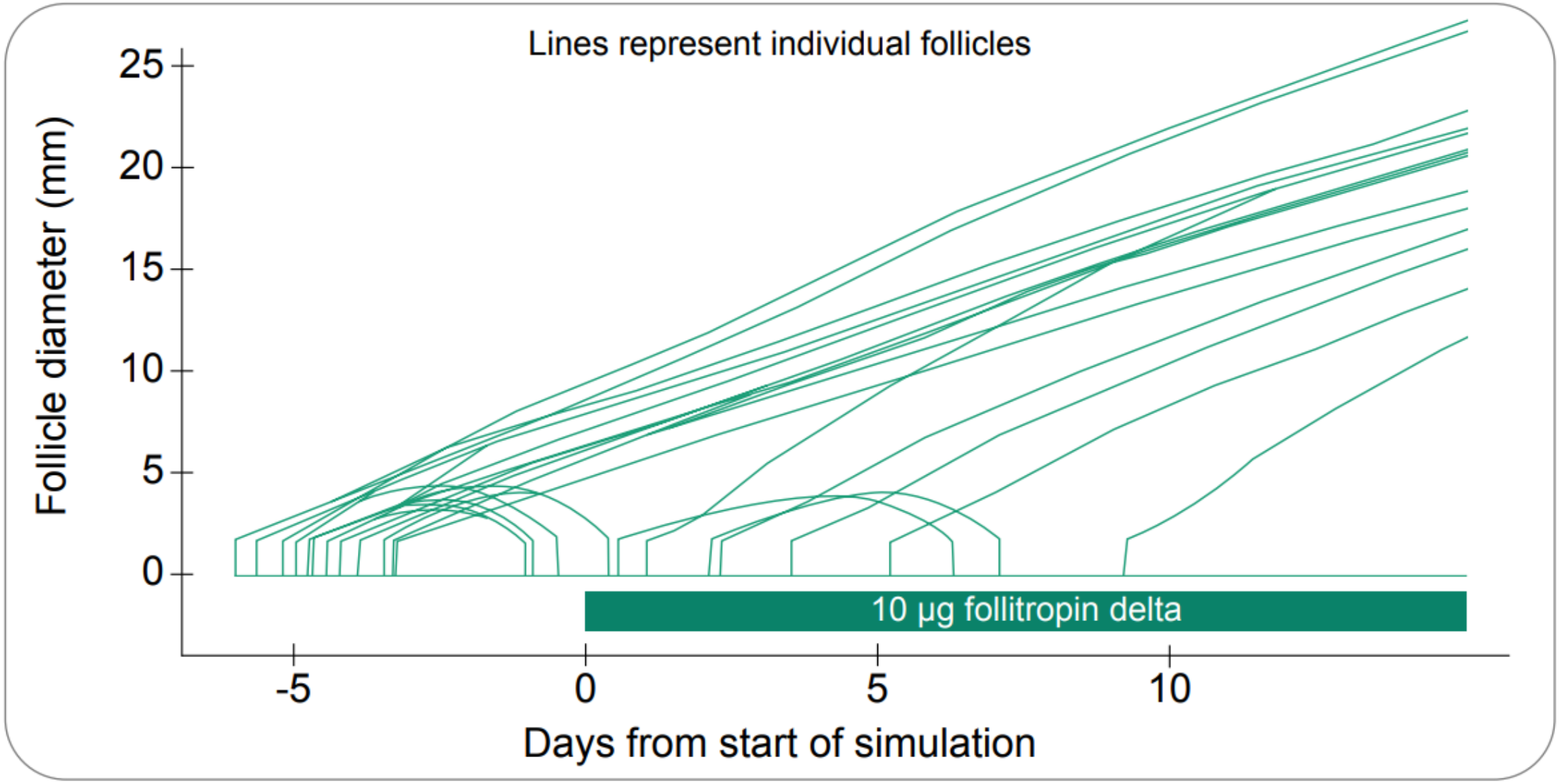
- Cellular level
  - Steroidogenesis, receptor dynamics in theca cells and granulosa cells
- Organ level: ovaries
  - Follicle numbers and size
- Organism
  - Pharmacokinetics, pituitary feedback

A complex model
- 70 model compartments
- 500+ model parameters
- 1700+ reactions

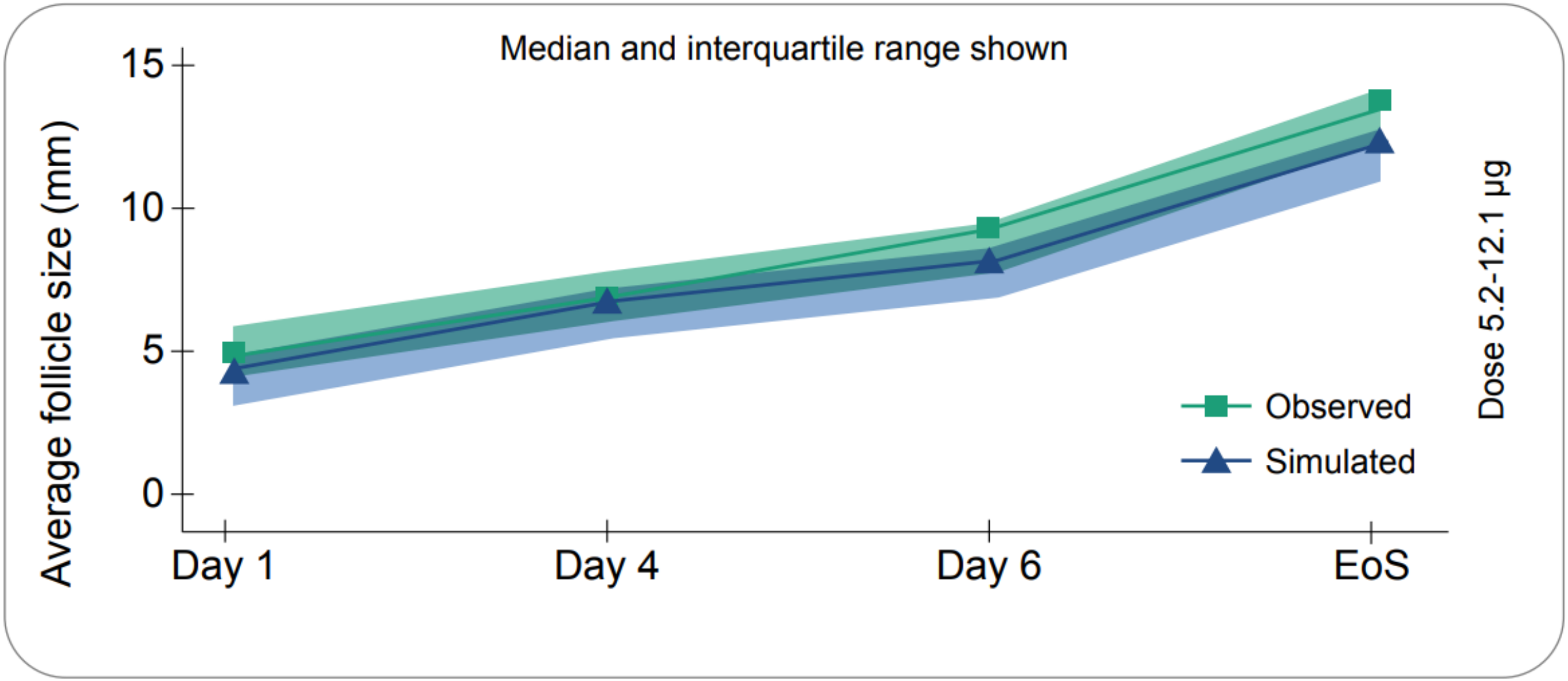


AMH, anti-Müllerian hormone; AR, androgen receptor; A_4, androstenedione; CYP17, cytochrome P450 17A1; CYP19, cytochrome P450 19A1 (aromatase); E_2, oestradiol; FSH, follicle-stimulating hormone; FSHR, follicle-stimulating hormone receptor; InhB, inhibin B; LH, luteinizing hormone; LHR, luteinizing hormone receptor; OS, ovarian stimulation; T, testosterone.

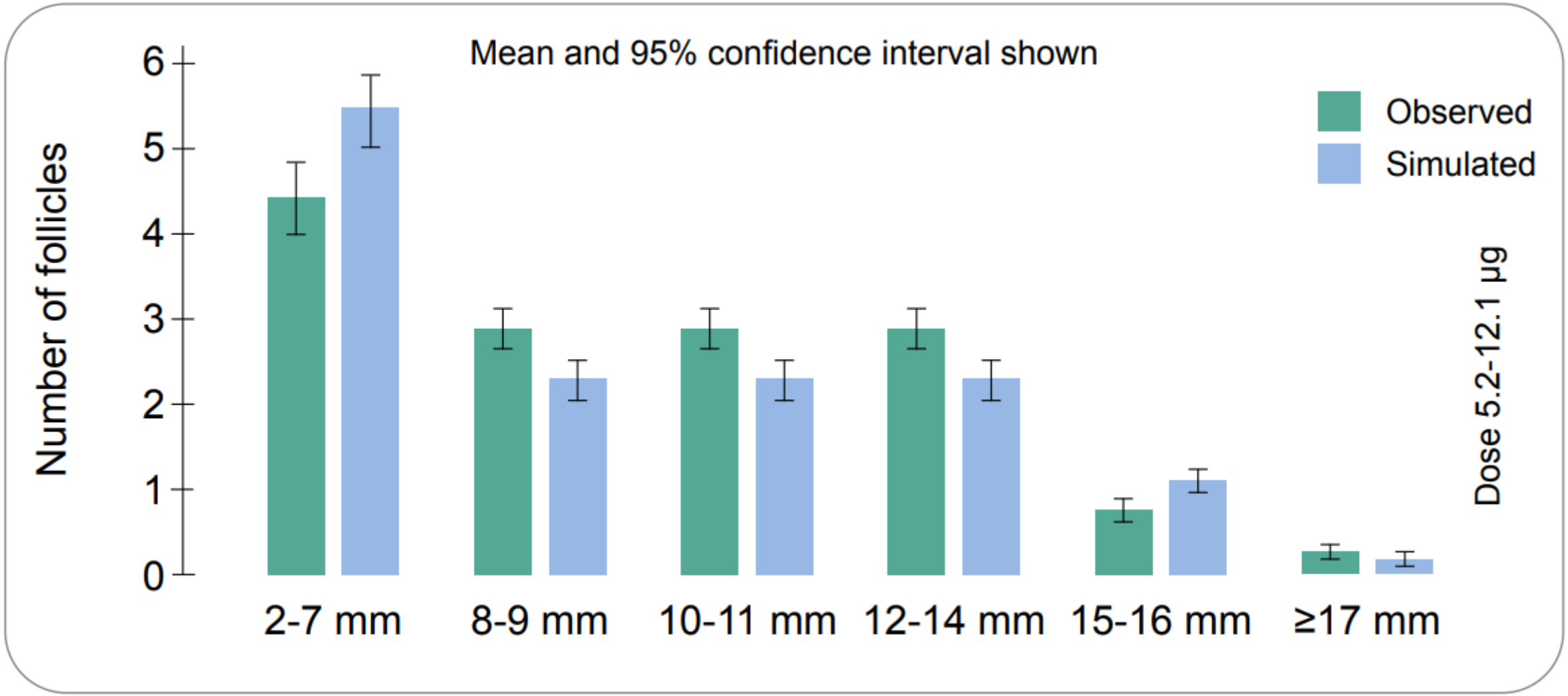
Simulation of follicle growth in a single patient



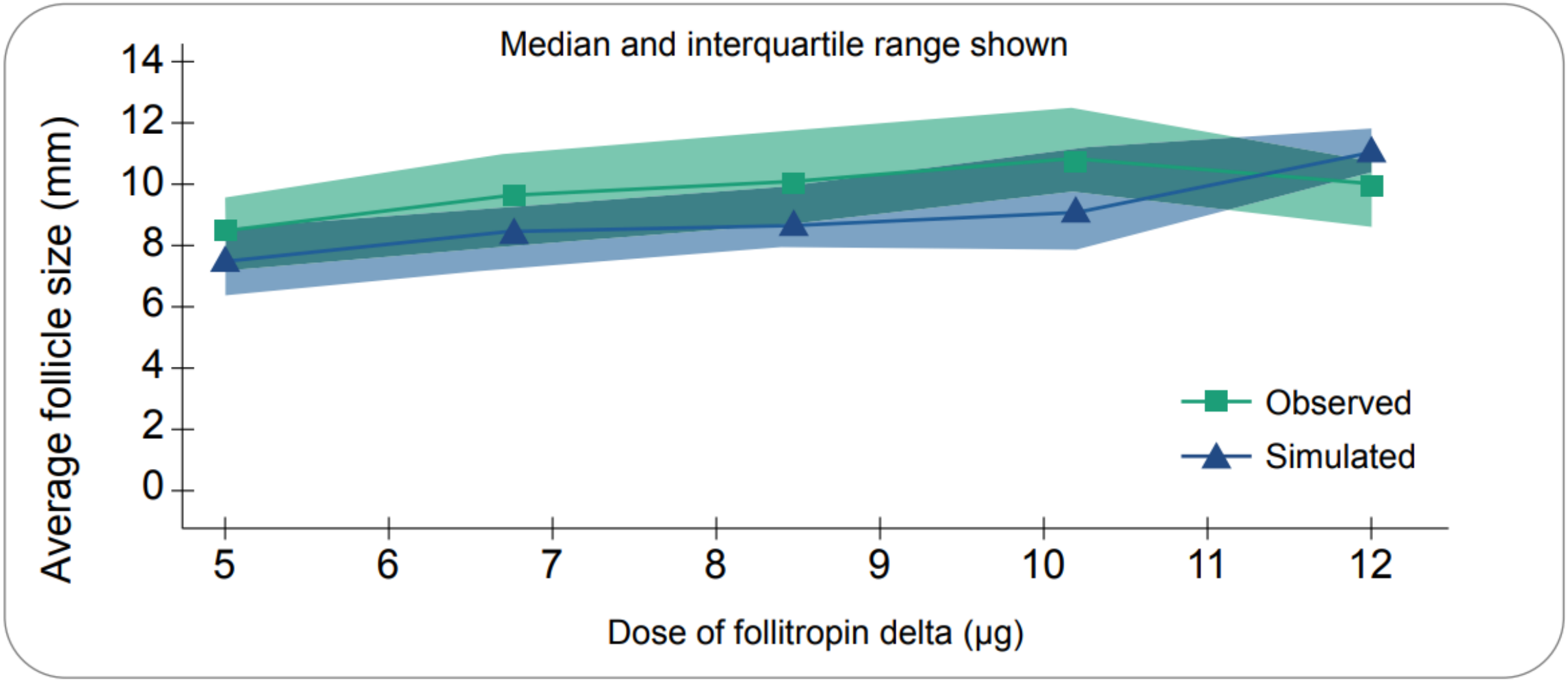
Observed and simulated follicle growth over time



Follicle size distributions on Day 6



Dose-dependence of follicle size on Day 6



FERRING PHARMACEUTICALS

Applications of the in silico model

Understand and predict how the ovary responds to drug interventions.

A large amount of information from different sources was integrated, clarifying biological connections and their relevance.

Ability to evaluate outcomes for many trial scenarios within a short time frame.

Hypothesis-generating through virtual studies that could not readily be performed with patients.

Inform dosing regimens and inclusion/ exclusion criteria for clinical trials.

Evaluating effects of changing e.g. dosing frequency or effects of dose changes during stimulation in various patient phenotypes.

Individualised forecasts of IVF outcome (potentially).

Creation of digital twins: instances of the model each with a unique set of parameters such that the model reflects a specific patient.

FERRING PHARMACEUTICALS

The team behind this work






Marcelo Behar, Alina Sode, Zhongyu Wang, Ruth Carcillo, Nikhil Patidar and Lars Aarby.





Christian Secchi, Erica Schoeller, Sarah Grover, Lars-Erik Kyhl and Pernille Maria Manuel.

Thank You. Questions?

FERRING PHARMACEUTICALS

Break

20 minutes

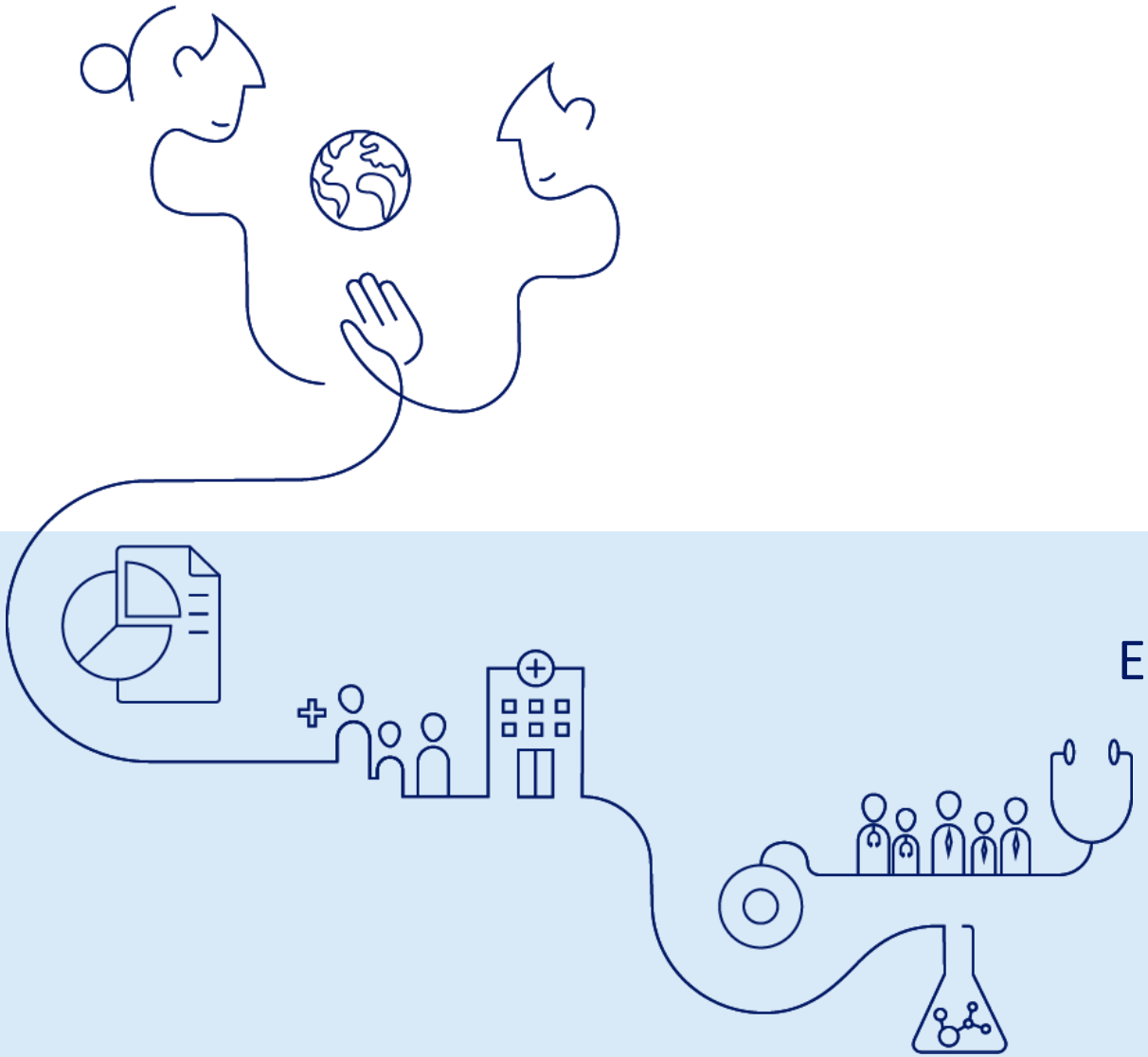
Today's programme

Timeslot	Speaker	Title
8:30 – 9:00		Breakfast and arrival
Programme starts:		
9:00-9:10	Randi (DSBS) Jonathan (FMS)	Welcome
Session 1:		**Topics in group sequential designs**
9:10 – 9:45	Corine Baayen (Ferring)	Design and analysis of group sequential trials for repeated measurements when pipeline data occurs: a comparison of methods
9:45 – 10:20	Henrik Thomsen (Novo)	Family-wise error for multiple time-to-event endpoints in a group sequential design
Break **(30 minutes)**		
Session 2:		**Working as a pharmaceutical statistician**
10:50 – 12:20	Anna Berglind (Novo) Jonas Häggström (Cytel) Niklas Berglind (AstraZeneca)	Medical statistics in practice – different ways of making a difference
Lunch **12:20-13:30**		

Timeslot	Speaker	Title
Session 3:		**Utilization of historical data**
13:30 – 14:05	Martin Bøg (Novo)	Historical Borrowing
14:05-14:40	Daniel Jonker (Ferring)	Advancing Precision Medicine with Innovative In Silico Approaches in Reproductive Medicine
Break **(20 minutes)**		
Session 4:		**Next Generation of young statisticians**
15:00-15:35 **On Teams**	Emilie Højbjerre-Frandsen (Novo & AAU, Ph.d. Berkeley US)	Prognostic Score Adjustment
15:35-16:00	Wrap up	
16:00		End of the day

Session 4: **Next Generation of young statisticians**

Session lead: Ketil



Prognostic score adjustment

Enhancing study power through historical data

Emilie Højbjerre-Frandsen, Industrial Ph.D. student at Novo Nordisk and AAU

Disclaimer

- Presenter is an employee of Novo Nordisk A/S
- Views and opinions expressed are those of the presenter and not necessarily Novo Nordisk A/S

- Work in the setup of an RCT
- The estimand of interest is the average treatment effect $\Psi = E[Y(1) - Y(0)]$
- Schuler A et al. *Increasing the efficiency of randomized trial estimates via linear adjustment for a prognostic score.* The International Journal of Biostatistics. 2021

A tutorial on improving RCT power using prognostic score adjustment

Emilie Højbjerre-Frandsen*[1,2] | Mathias Lerbech Jeppesen & Rasmus Kuhr Jensen[1] | Claus Dethlefsen[1,2] | Rasmus Waagepetersen[2]

[1]Biostatistics, Novo Nordisk A/S, Vandtårnsvej 114, Søborg, Denmark
[2]Department of Mathematical Sciences, Aalborg University, Skjernvej 4A, Aalborg Øst, Denmark

Correspondence
*Emilie Højbjerre-Frandsen, Aalborg, Denmark. Email: ehfd@novonordisk.com

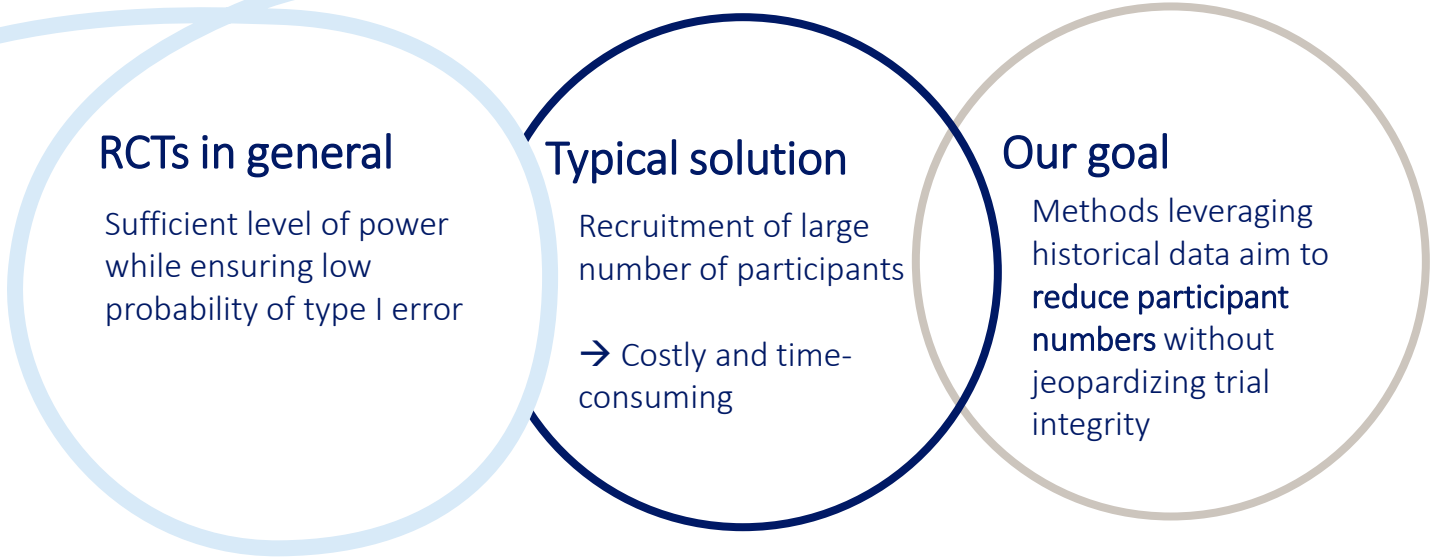
Abstract

The use of historical data to increase power in clinical trials has been a topic of interest for many years. A recent approach adjusts linearly for a prognostic score. This is supported by asymptotic results involving influence functions for asymptotically linear estimators. We provide further justification by a finite sample optimality result. A simulation study is conducted to investigate the performance in finite samples, comparing to standard procedures such as propensity score matching for RCTs (PSM-RCT) and ANCOVA using simple baseline adjustment. The simulation study investigates four different data generating scenarios to test the performance and sensitivity of the method under different assumptions. Unlike PSM-RCT, linear adjustment for a prognostic score avoids biased treatment effect estimates and maintains control of type I error probability. The simulation study shows that the method is robust against deviations from method assumptions and poor performance of the

*Randomised clinical trials (RCT)

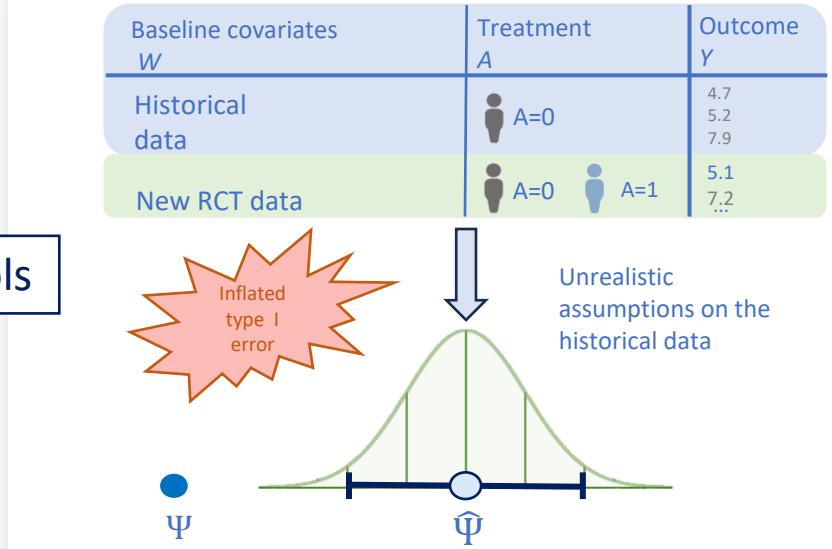
*https://github.com/NNpackages/PostCard

Motivation

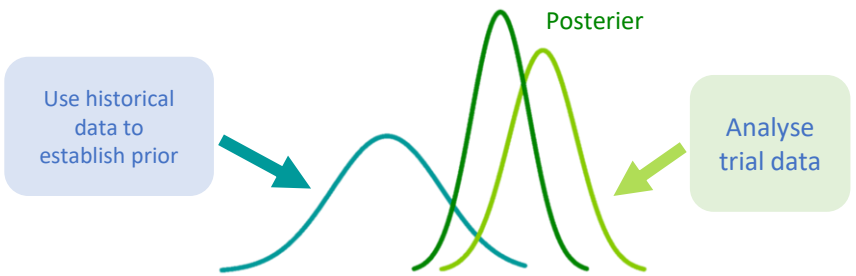

RCTs in general
Sufficient level of power while ensuring low probability of type I error
Typical solution
Recruitment of large number of participants
→ Costly and time-consuming
Our goal
Methods leveraging historical data aim to **reduce participant numbers** without jeopardizing trial integrity

Existing solutions

External controls



Baysian statistics



Proposed solution

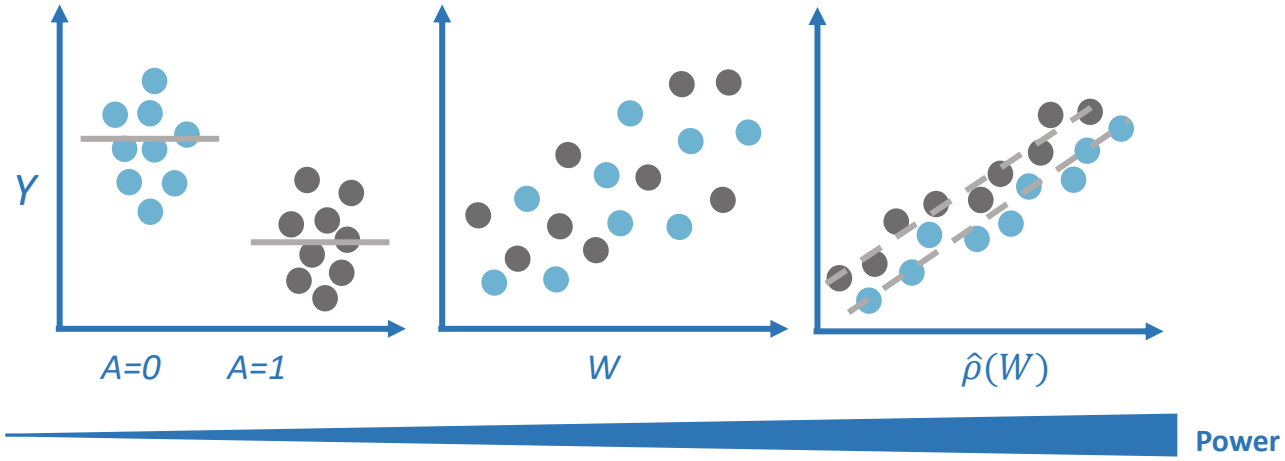
- Determine a prognostic score estimated from historical data

$$\hat{\rho}(W) = \hat{E}[Y \mid W, A = 0]$$

- Use ANCOVA model adjusting for $\hat{\rho}(W)$

$$Y = ATE \cdot A + \beta \cdot W + \alpha \cdot \hat{\rho}(W) + error$$

- The higher correlation with the outcome the higher power increase



How the method works

Step 1

- Curate historical data from different sources
- Train a prognostic model $\hat{\rho}$

Step 2

- Evaluate the performance of the prognostic model
- **Correlation between outcomes and predicted outcomes** on an independent test data set

Step 3

- Predict the prognostic scores for each of the participants in the new trial

Patient Number	W_1	W_2	W_3	...	W_p	A	Y
1	M	48	175		55	1	34
2	M	34	179		64	0	42
3	K	18	189		87	1	67
4	M	22	165		35	0	21
....							

Patient Number	W_1	W_2	W_3	...	W_p	A	Y	$\hat{\rho}$
1	M	48	175		55	1	34	$\hat{\rho}(w^1)$
2	M	34	179		64	0	42	$\hat{\rho}(w^2)$
3	K	18	189		87	1	67	$\hat{\rho}(w^3)$
4	M	22	165		35	0	21	$\hat{\rho}(w^4)$
....								

Step 4

- Use ANCOVA model adjusting for $\hat{\rho}(W)$.
- Type I error control
- Under specific requirements $\widehat{ATE}_{\hat{\rho}}$ has the lowest possible asymptotic variance among RAL estimators

Theorem 1

Assume that

$$E[Y(1)|\,W] = E[Y(0)|W] + ATE.$$

Also assume that the conditional variance

$$Var(Y|A,W) = \sigma^2$$

does not depend on (*A, W*). Then the OLS estimate of the ATE obtained from an ANCOVA model with design matrix $X = [A \quad E[Y(0)|\,W]]$ is an unbiased estimator of the ATE and has the lowest possible variance among all estimators of the ATE that are conditionally unbiased given (*A, W*) and of the linear form

$$B(W,A)Y$$

where the $1 \times n$ matrix $B(W,A)$ is a function of *W* and *A*.

**Regular and asymptotically linear (RAL)*

Simulation study

Data simulation and scenarios

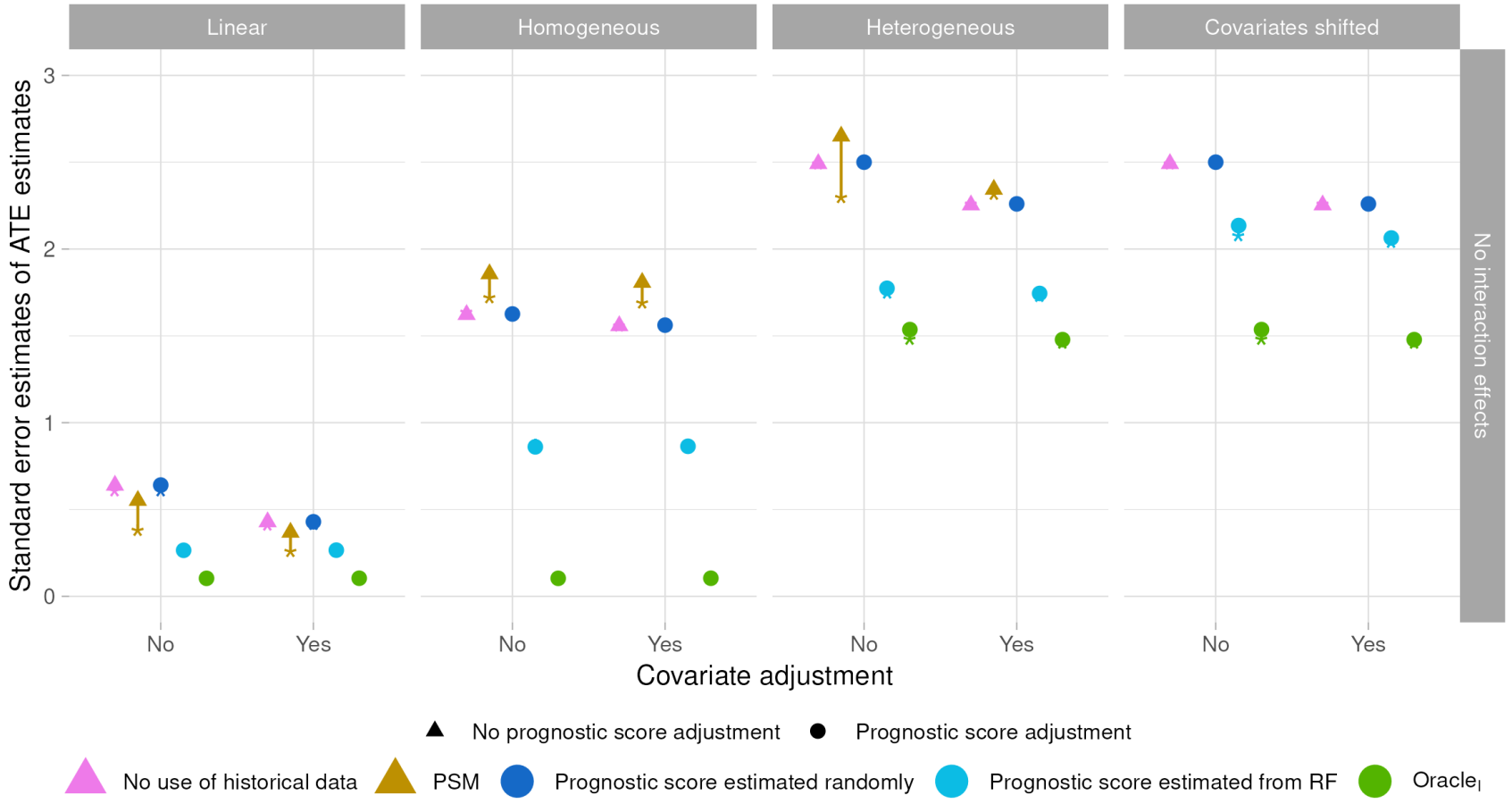
- We perform a simulation study to test the **finite sample properties** and **sensitivity to method assumptions**
- Data is simulated conditional on W and A from a normal distribution
- $$Y(A) \mid W \sim \aleph\left(\boldsymbol{a}^{\boldsymbol{T}} W : W + \boldsymbol{b}^{\boldsymbol{T}} W + \boldsymbol{c}^{\boldsymbol{T}} WA + ATE \cdot A, \sigma^2\right) \quad (1)$$

Scenario	a	b	c	d
Linear covariate effects	0	1	0	0
Homogeneous treatment effect	0.5	1	0	0
Heterogeneous treatment effect	0.5	1	4	0
Covariates shifted	0.5	1	4	4

Table 1. Coefficients of equation (1) for four data generation scenarios. a: degree of non-linearity and interaction effects. b: linear main effects. c: interaction effect with covariates and the treatment. d: is the mean of the normal distribution that the covariates are generated from in the historical data.

Error estimates

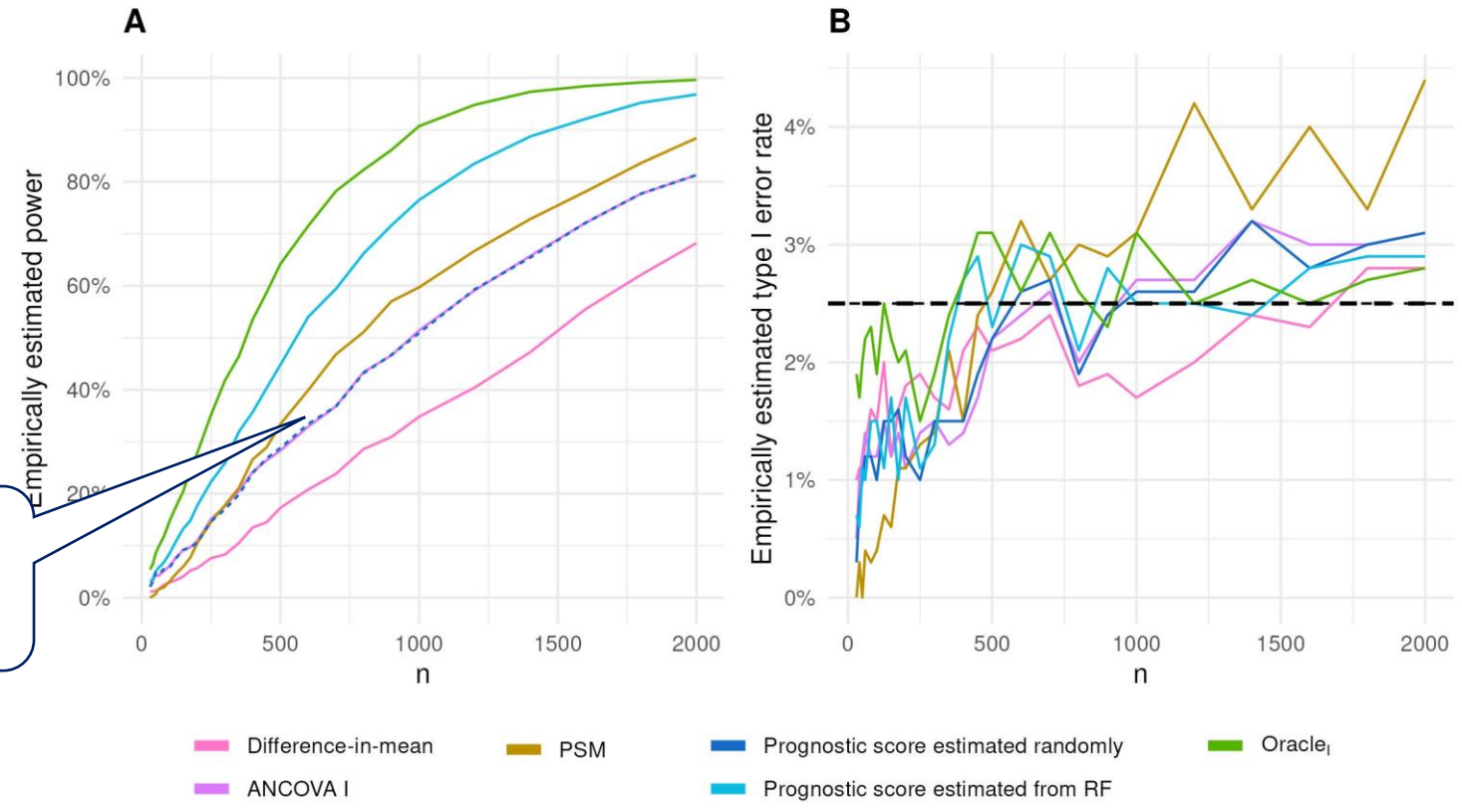
Filled points are mean of standard error estimates. Crosses are the root mean squared error (RMSE).



*Propensity score matching (PSM)
*** Random and Random forest refers two the prognostic model being used to determine the predicted outcomes for each participant and afterwards adjusted for

Power and type I error

Robustness property



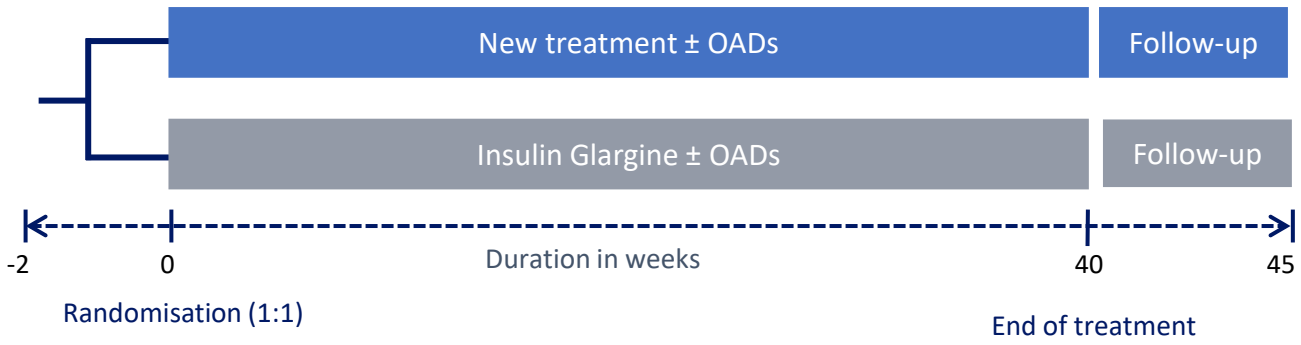
*Propensity score matching (PSM)
** n is the sample size of the current RCT data, with the historical data amount being n'=5*n
*** Random and Random forest refers two the prognostic model being used to determine the predicted outcomes for each participant and afterwards adjusted for

Case study

Trial design

474 patients

- T2D
- HbA1c ≥ 8.0%
- BMI ≤ 40 kg/m^2



Study information

- Open label, Parallel group
- Treat-to-target
  - 3.9 – 5.0 mmol/L

Study objective

To confirm the efficacy (superiority on HbA_{1c}) and compare safety of new treatment compared with daily insulin glargine, with or without OADs in participants with T2D

Study Estimand

Primary: The treatment effect between new treatment and daily insulin glargine in change in HbA_{1c} from baseline to week 40 in participants with T2D regardless of discontinuation of randomised treatment for any reason and regardless of initiation of non-randomised insulin treatment or additional anti-diabetic treatments for more than 2 weeks

Key endpoints

- **Primary:** Change in HbA_{1c} from baseline to week 40
- **Secondary**:
  - Change in body weight from baseline to week 40
  - CGM based endpoints from week 36 to week 40**
  - Number of level 2 and 3 hypoglycaemic episodes from baseline to week 45

**Type 2 diabetes (T2D), other antidiabetic drug (OAD), continuous glucose monitoring (CGM)*

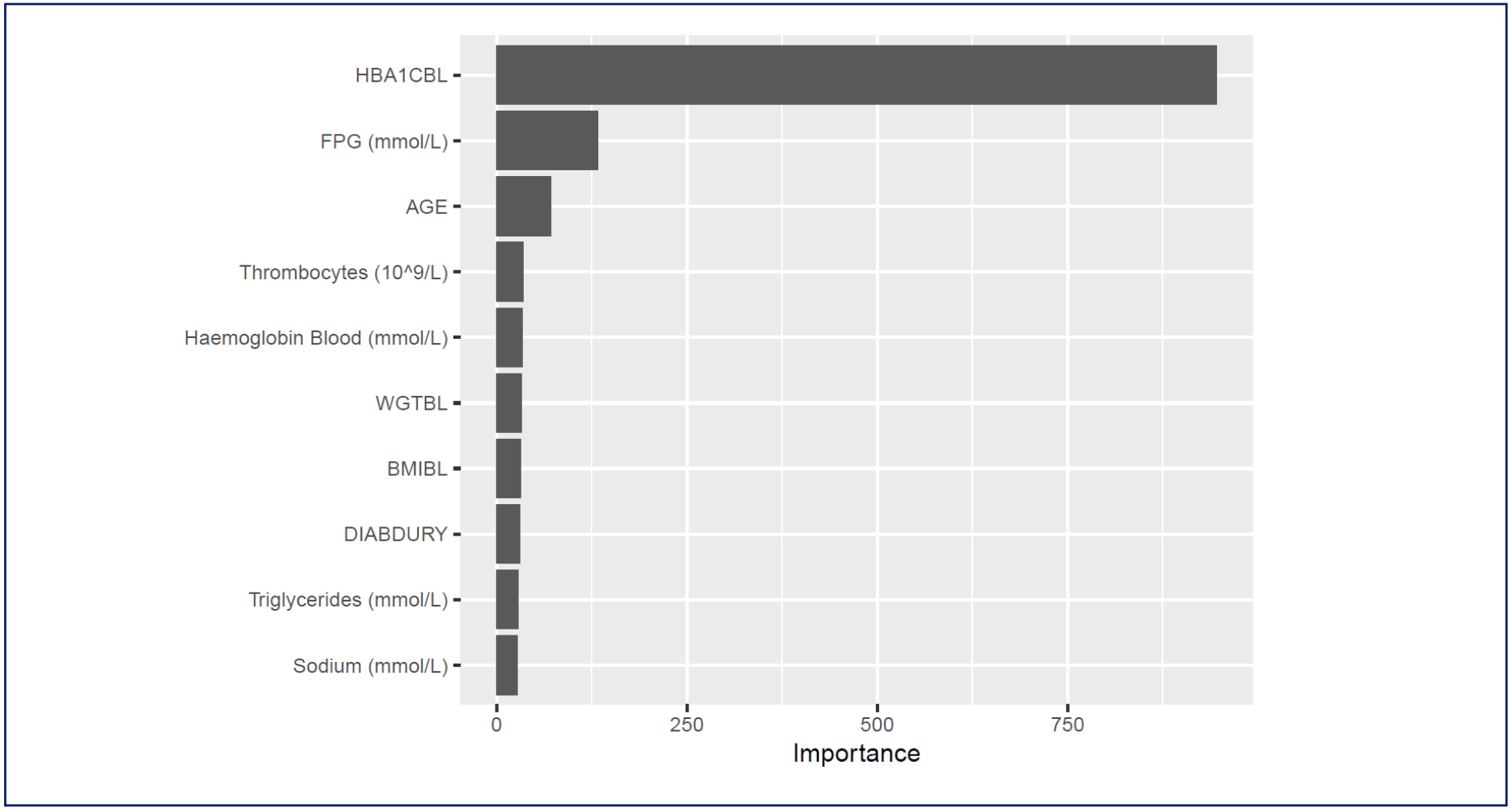
Phase 3b

Why prognostic score adjustment?

- Label expansion
  - Crucial to have study results ready between approval and launch

- **Solution:** Minimise the number participants in the study by leveraging historical data, while maintaining power and without compromising the type I error rate

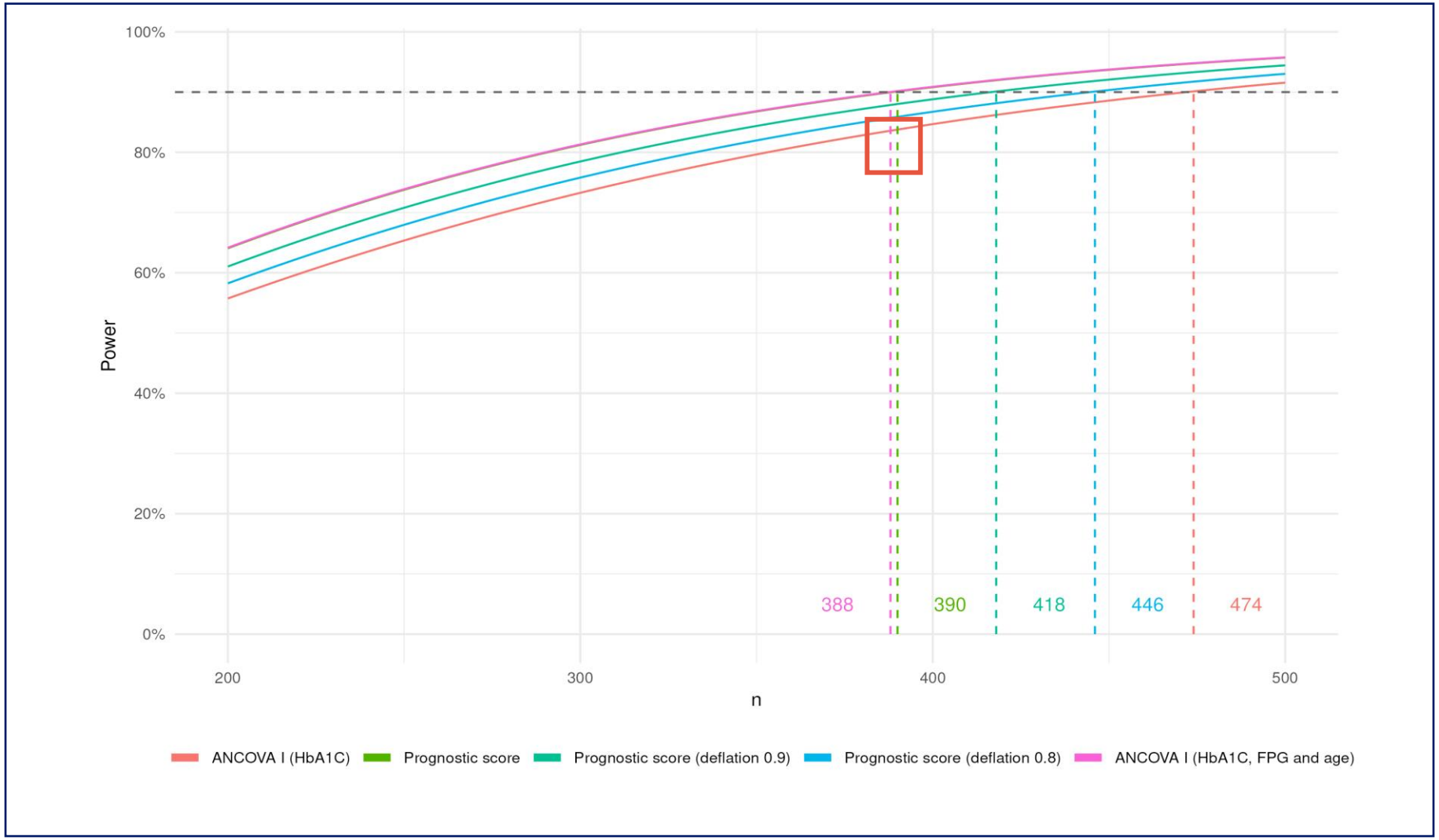
Variable importance



VIMP plot from Random forest machine learning model.

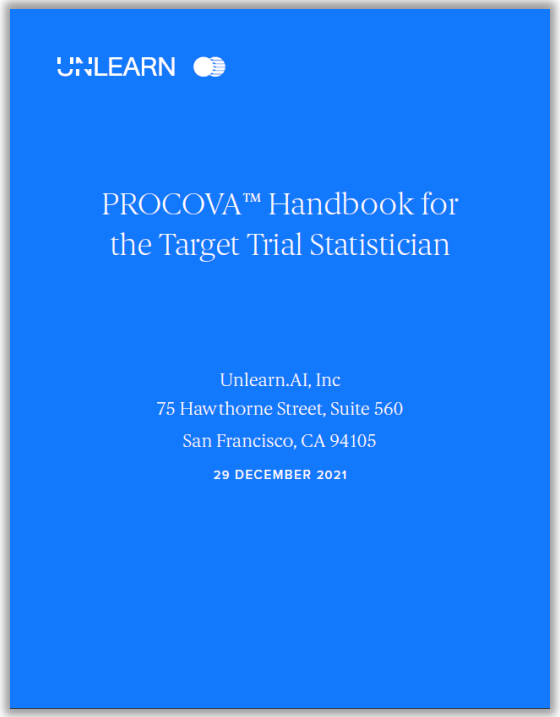
The variable importance measure is computed from permuting out-of-bag (OOB) data; for each tree, the prediction error on the OOB portion of the data is recorded (error rate for classification and MSE for regression). Then the same is done after permuting each predictor variable. The difference between the two are then averaged over all trees in the forest and normalized by the standard deviation of the differences. If the standard deviation of the differences is equal to 0 for a variable, the division is not done (but the average is almost always equal to 0 in that case).

Required sample size



Practical experience

- Difficult to combine historical data into a curated data set that can be used for model fitting
- Strong predictors such as baseline HbA1c may limit the gain in precision if already included in the analysis as covariates
- Choice of deflation parameter could seem arbitrary
  - Guidance can be found in *PROCOVA™ Handbook for the Target Trial Statistician**



* procovatm-handbook_en.pdf (europa.eu)

Compromises

- Lack of power if prognostic score has a lower effect than assumed
  - However, not lower than the power for the analysis without prognostic score with reduced number of participants
- Smaller sample size gives lower power for the statistical analyses of other endpoints that does not have prognostic score adjustment
- Subgroup analyses cannot be done using prognostic score adjustment since the effect could already be captured through the prognostic model
- Is only for continuous endpoints

Questions?

Wrap up and closing

Session lead: Carl-Fredrik