Statistics Boost 21

Trial Success Probabilities – From Power to DDCP

Godwin Yung, PhD

PDD Methods, Collaboration, and Outreach Group (PDD MCO)

Statistics Boost *Standard* Series (since 2022)

To **upskill the entire organization** on **statistical methodology relevant to drug development**.

go.roche.com/statsboost: past recordings and subscription information

Upcoming sessions (1h/month)
Survival analysis – non-proportional hazards
Statistics in early development and dose finding
Statistics in early oncology development
Futility interim analyses: Why, how, when?
Efficacy interim analyses: Why, how, when?

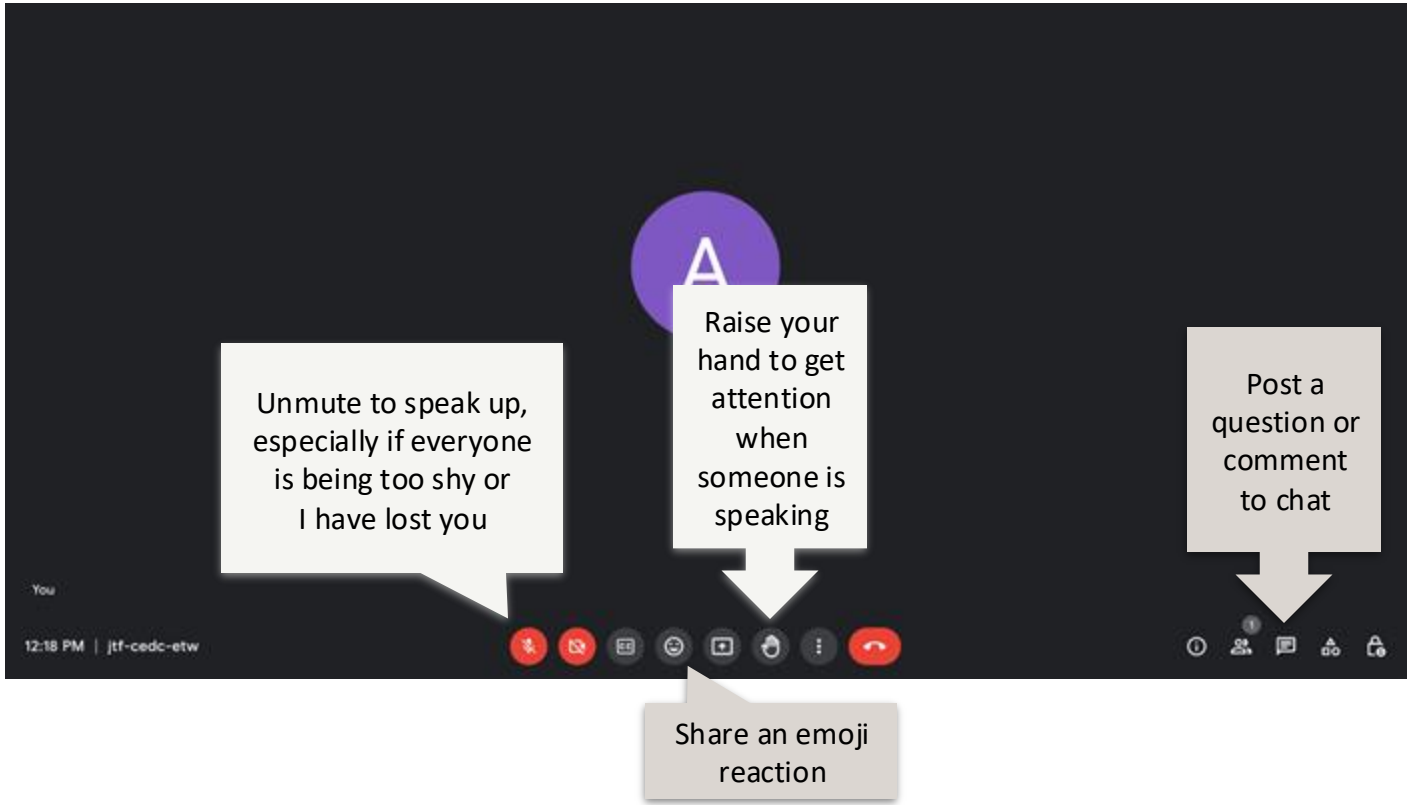
Statistics Boost *Advanced* (New)

To **upskill statisticians** on **statistical methodology relevant to drug development.**

PD connect post with all infos

Upcoming sessions (1h/month)
Group-sequential trials - Inference after stopping and hierarchical testing (Schedule for 2025 to be announced)

Interacting with each other



Thank you!

Members of MCO

Kaspar Rufibach (for previous thinking)

Ulrich Beyer

Bernadette Surujbally

Maoxia Zheng

Chenglin Ye

Jessie Hsu

Today's program

"If our Ph3 studies are powered at 80%, why aren't we seeing 80% success rates?"

- What does it mean to power a study at 80%?
- If a study is powered at 80%, what are the chances of it turning out positive?
- How can we increase Ph3 PTS?

Today's program

"If our Ph3 studies are powered at 80%, why aren't we seeing 80% success rates?"

- What does it mean to power a study at 80%?
- If a study is powered at 80%, what are the chances of it turning out positive?
- How can we increase Ph3 PTS?

Power

"How often do we claim drug works if in population it does?"

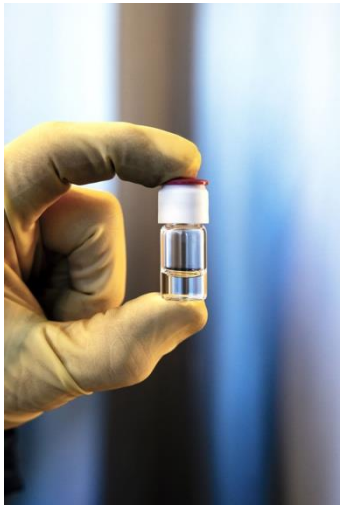
↓

"The probability of observing an estimated treatment effect that is statistically significant… given a particular true treatment effect."

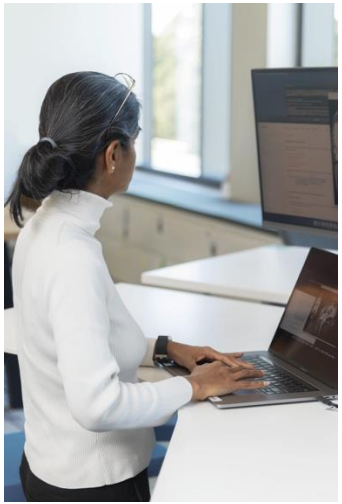
Why do we conduct clinical trials?

"Does drug A work in population X?"

Trial
Experimentation



Data



Inferential Statistics

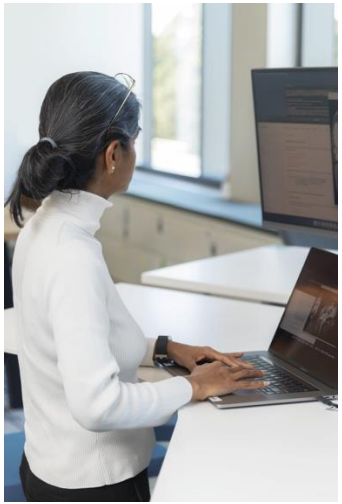
(probabilistic statements involving uncertainty/variability)

Why do we conduct clinical trials?

"Does drug A work in population X?"

Trial
Experimentation

Data



Study design:
Enables inference by asking, "If [condition] and we do [action], then [probabilistic result]."

GYM329 in obesity

GYM329 as an add-on therapy to incretins

Proposed Ph3 trial design[1]

- Efficacy analysis: Percent body weight loss at 1y
- Sample size: 104 patients[2], randomized 1-to-1
- Significance level: one-sided 2.5%
- "Study is **powered at 80%** to detect a **5.0% loss**"
- "An observed treatment effect of **3.5% loss** or more may be deemed statistically significant", i.e., the **minimum detectable difference (MDD)**

PROTOCOL

TITLE:

PROTOCOL NUMBER:

TRIAL NAME: [optional]

VERSION NUMBER:

TEST PRODUCT:

TRIAL PHASE:

SPONSOR NAME AND LEGAL REGISTERED ADDRESS:

SHORT TITLE: [optional]

REGULATORY AGENCY IDENTIFIER NUMBERS: IND Number: EU CT Number: CIV ID: PIS ID: NCT Number:

APPROVAL: Refer to electronic signature and date stamp on the first page of this document

MEDICAL MONITOR CONTACT:

SERIOUS ADVERSE EVENT REPORTING METHOD: For detailed reporting of serious adverse events to the Sponsor, refer to Section 9.2.

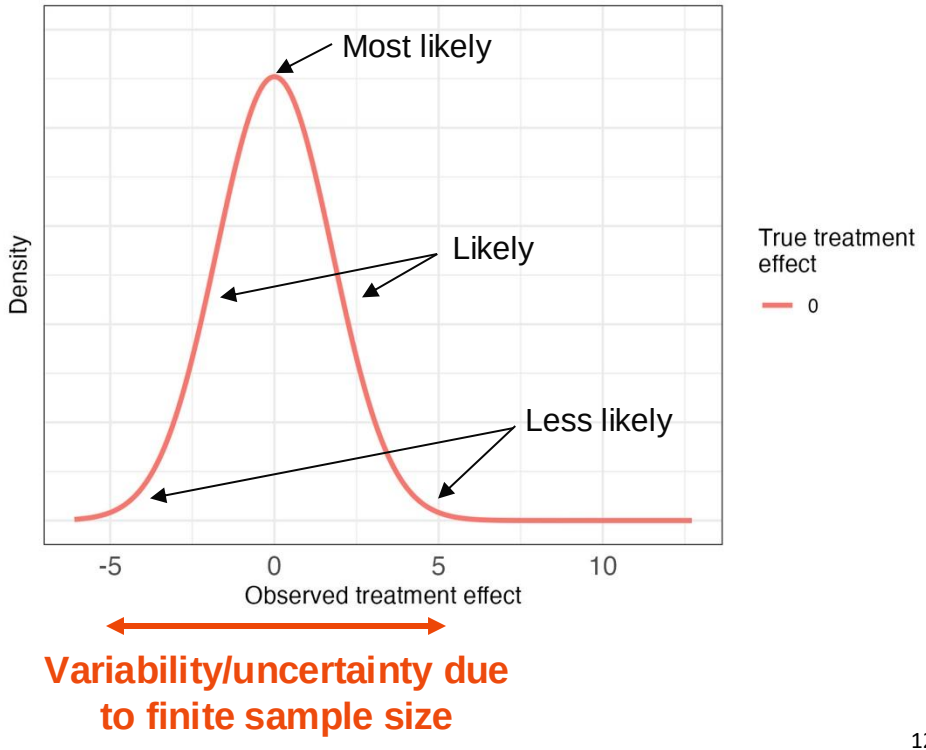
OneRoche Protocol Template
Version 1, 31 January 2024

[1]Modified for the purpose of this presentation

[2]Assumes a standard deviation of 9% in both arms

What Ph3 treatment effect might we observe under the null hypothesis of ineffective treatment?

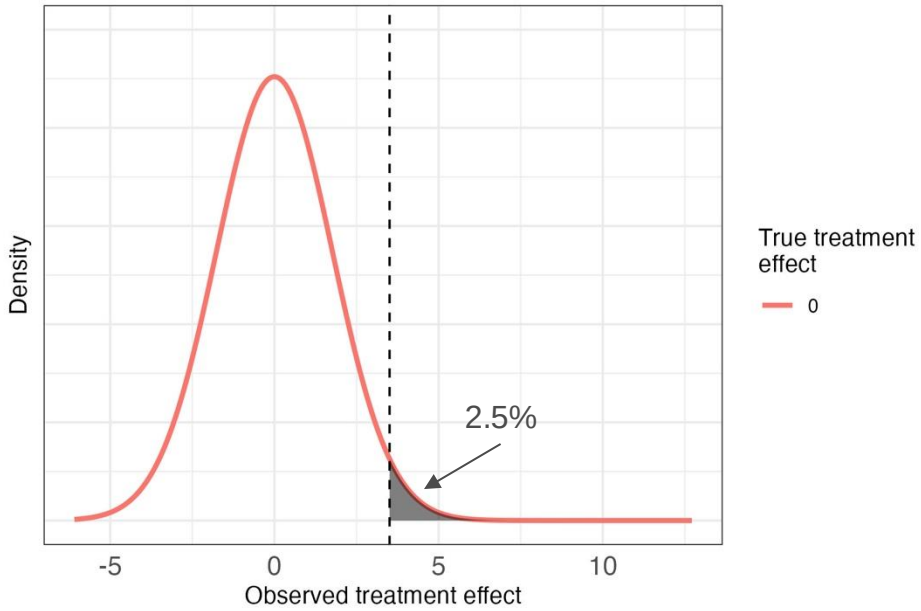
- Sample size: 104 patients
- True treatment effect = 0.0% loss



Critical value

Result for which trial will just be statistically significant

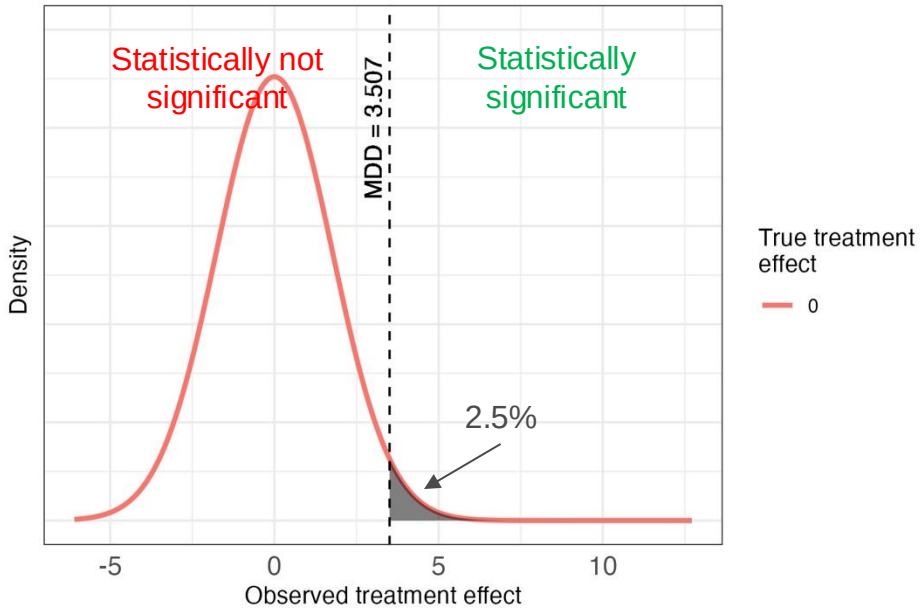
- **Significance level:** critical value on the p-value scale (e.g., one-sided 2.5%)
  - False positive rate: “If, applied to the entire patient population, incretin+GYM329 results in **0.0% body weight loss at 1y**, then our Ph3 trial has a 2.5% chance of achieving statistical significance”
  - Driven by health authority



Critical value

Result for which trial will just be statistically significant

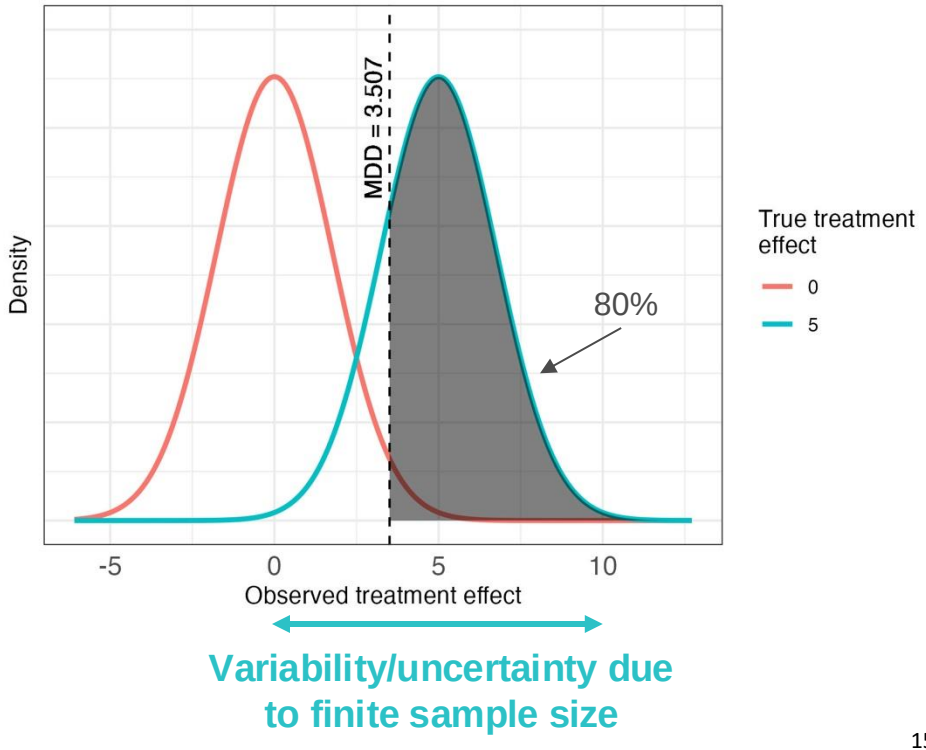
- **Minimal detectable difference (MDD):** critical value on the treatment effect scale (e.g., 3.5% loss)
  - Recommend aligning with MVP
  - "So what?" scenario



Power

Probability of observing an estimated treatment effect that is statistically significant (i.e., meets the MDD) … **given a particular true treatment effect**

"**If**, applied to the entire patient population, incretin+GYM329 results in **5.0% body weight loss at 1y**, then our Ph3 trial has an 80% chance of achieving statistical significance (i.e., observing a treatment effect > 3.5% loss)"



Today's program

"If our Ph3 studies are powered at 80%, why aren't we seeing 80% success rates?"

- What does it mean to power a study at 80%?
- If a study is powered at 80%, what are the chances of it turning out positive?
- How can we increase Ph3 PTS?

Power

Probability of observing an estimated treatment effect that is statistically significant (i.e., meets the MDD) … **given a particular true treatment effect**

"**If**, applied to the entire patient population, incretin+GYM329 results in **5.0% body weight loss at 1y**, then our Ph3 trial has an 80% chance of achieving statistical significance (i.e., observing a treatment effect > 3.5% loss)"

But we don't know what the true treatment effect is!

What if the true treatment effect is 3.0% loss? Or 7.0%?

Power depends on sample size *and* true treatment effect

From Power to DDCP

"Power is the probability of observing an estimated treatment effect that is statistically significant… **given a particular true treatment effect**."

"Data driven conditional probability (DDCP) is the probability of observing an estimated treatment effect that is statistically significant… **given what we know about the true treatment effect**."

DDCP acknowledges additional uncertainty

"**Power** is the probability of observing an estimated treatment effect that is statistically significant… **given a particular true treatment effect**."

Power assumes that a particular treatment effect is the truth.

"**Data driven conditional probability (DDCP)** is the probability of observing an estimated treatment effect that is statistically significant… **given what we know about the true treatment effect**."

DDCP recognizes that we don't know what the true treatment effect is, but we might have an idea of the values it can take on and the likelihood of these values.

Toy Example

Power

% loss	Likelihood of being true*	Power if true
5.0	1.00	80%

*Fixed assumption

Probability of observing an estimated treatment effect that is statistically significant:

1.00 x 80% = 80%

DDCP

% loss	Likelihood of being true*	Power if true
3.0	0.25	39%
5.0	0.50	80%
7.0	0.25	98%

*Given observed data (data-driven)

Probability of observing an estimated treatment effect that is statistically significant:

0.25 x 39% + 0.50 x 80% + 0.25 x 98% = 74%

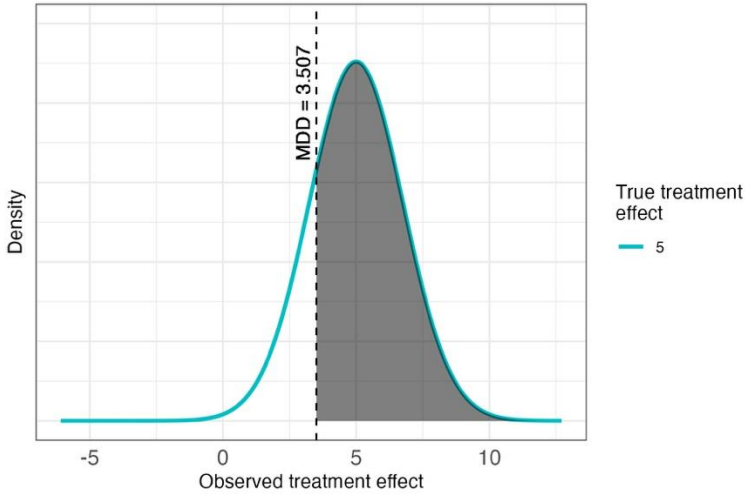
Toy Example - Power

% loss	Likelihood of being true*	Power if true
5.0	1.00	80%

*Fixed assumption

Probability of observing an estimated treatment effect that is statistically significant:

1.00 x 80% = 80%



Variability/uncertainty due to finite sample size

No uncertainty in the true treatment effect

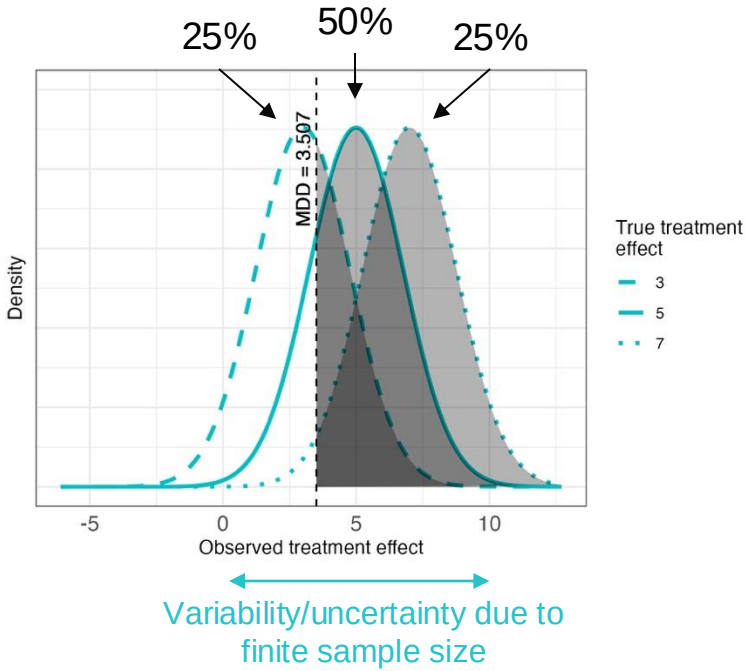
Toy Example - DDCP

% loss	Likelihood of being true*	Power if true
3.0	0.25	39%
5.0	0.50	80%
7.0	0.25	98%

*Given observed data (data-driven)

Probability of observing an estimated treatment effect that is statistically significant:

0.25 x 39% + 0.50 x 80% + 0.25 x 98% = 74%



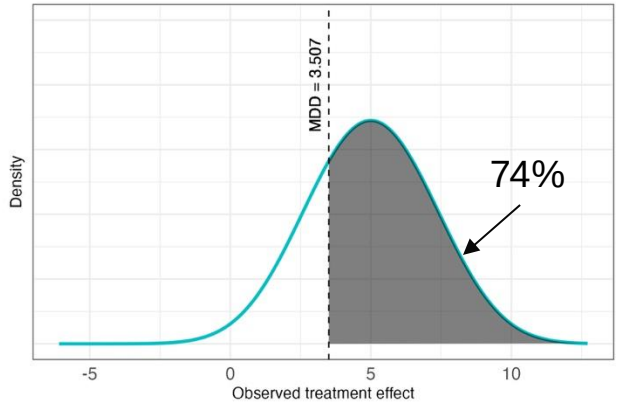
Toy Example - DDCP

% loss	Likelihood of being true*	Power if true
3.0	0.25	39%
5.0	0.50	80%
7.0	0.25	98%

*Given observed data (data-driven)

Probability of observing an estimated treatment effect that is statistically significant:

0.25 x 98% + 0.50 x 80% + 0.25 x 42% = 74%



Variability/uncertainty due to finite sample size

Variability/uncertainty due to finite sample size *and* multiple potential true treatment effects

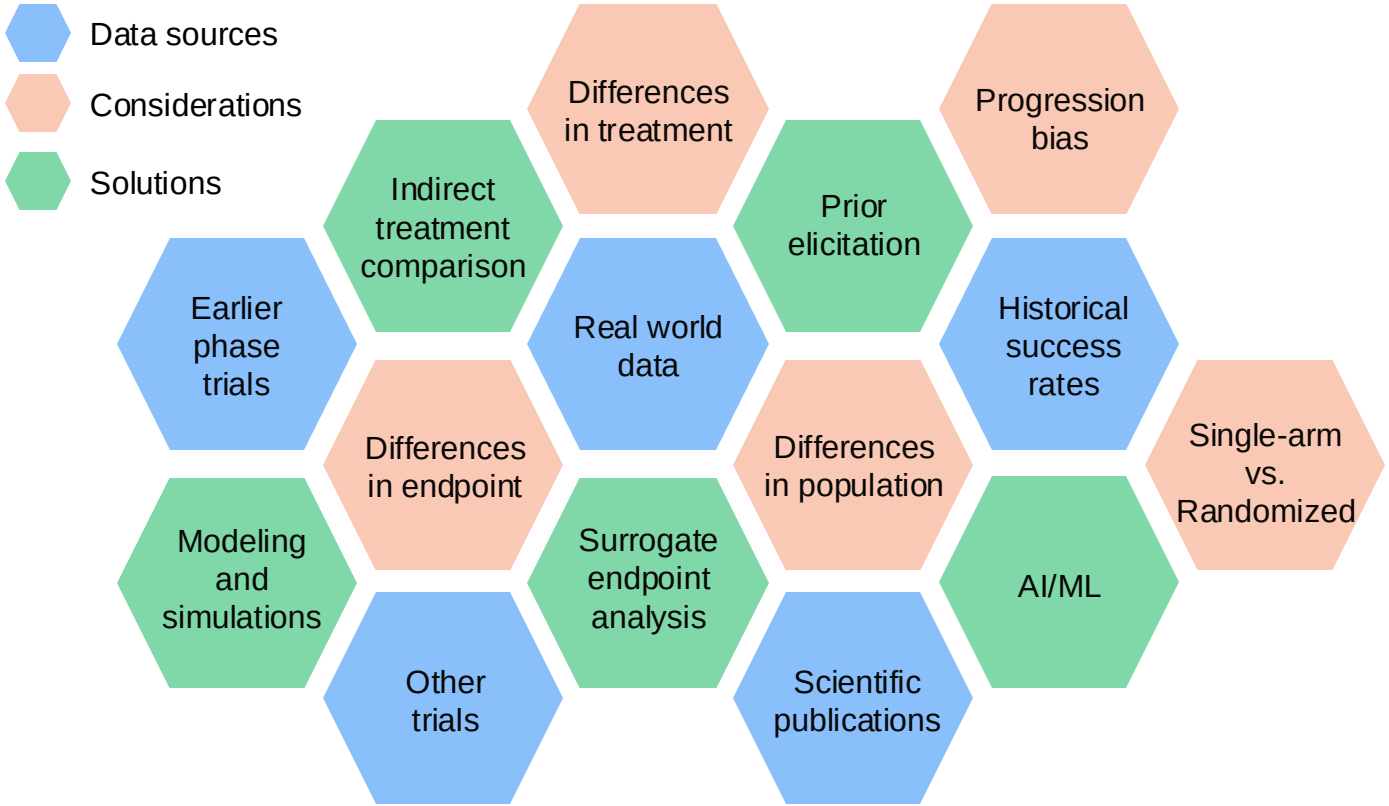
“DDCP is the probability of observing an estimated treatment effect that is statistically significant… **given what we know about the true treatment effect**.”

“**expected power** of a trial, taking into account various **potential values for the true treatment effect**”

“**best guess** on how likely a trial will actually be successful, based on **available info**”

How can we inform the “prior”?

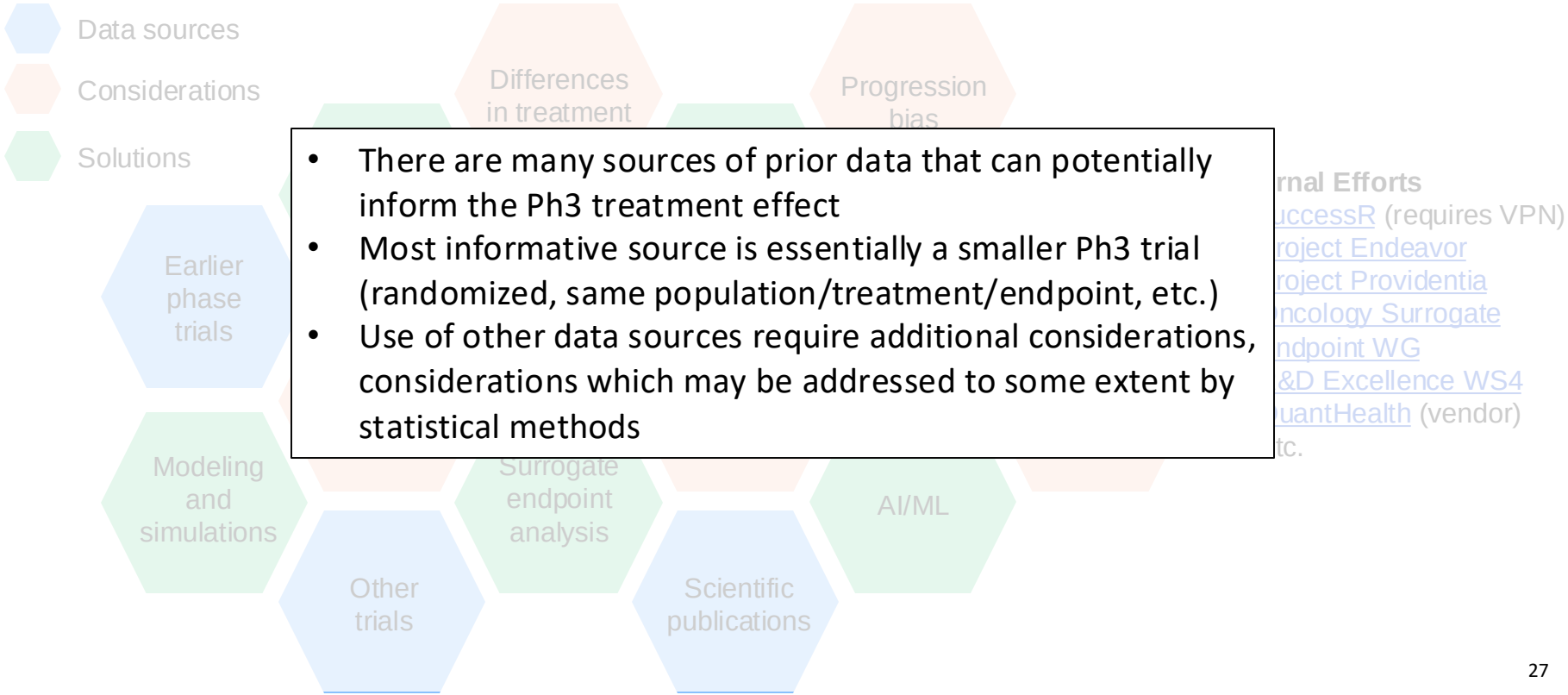
Informing the “prior”



Internal Efforts

- successR (requires VPN)
- Project Endeavor
- Project Providentia
- Oncology Surrogate Endpoint WG
- R&D Excellence WS4
- QuantHealth (vendor)
- Etc.

Informing the “prior”



- There are many sources of prior data that can potentially inform the Ph3 treatment effect
- Most informative source is essentially a smaller Ph3 trial (randomized, same population/treatment/endpoint, etc.)
- Use of other data sources require additional considerations, considerations which may be addressed to some extent by statistical methods

Since a trial's DDCP depends on the prior information, is there anything we can say about the general relationship between DDCP and Power?

Toy Example

Power

% loss	Likelihood of being true*	Power if true
5.0	1	

*Fixed assumption

Probability of observing an estimated treatment effect that is statistically significant:

1.00 x 80% = 80%

DDCP

% loss	Likelihood of being true*	Power if true
		39%
		80%
7.0	0.25	98%

*Given observed data (data-driven)

Probability of observing an estimated treatment effect that is statistically significant:

0.25 x 39% + 0.50 x 80% + 0.25 x 98% = 74%

The fact that DDCP < Power is not a coincidence!

Typically, DDCP < Power

- **Why, "The Homebuyer's Gamble":** Potential downside of a weaker treatment effect exceeds the potential upside of a stronger treatment effect.

- **Implication:** *Unless* your prior indicates a *strong belief* that the true treatment effect is *better* than the (single) treatment effect assumed for powering the trial, DDCP < Power!

Example

% loss	Likelihood of being true	Power if true
3.0	0.25	39%
5.0	0.50	80%
7.0	0.25	98%

Equally distant from 5.0

Equally likely

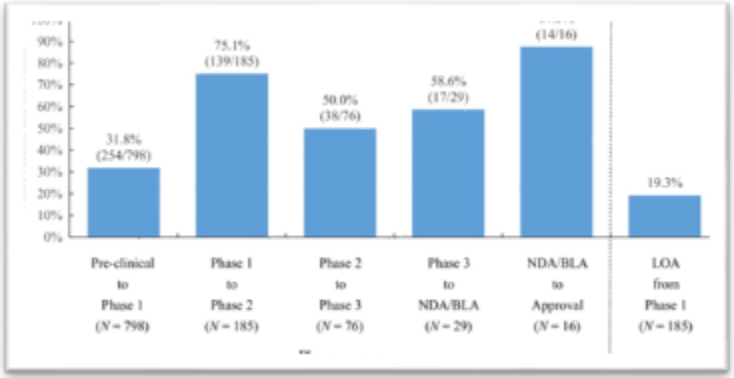
Differential impact!

Typically, DDCP < Power

- **This phenomenon helps to explain the industry-wide ~60% Ph3 success rate!**

- Companies that use DDCP (or assurance) for quantitative decision making: Novartis, GSK, J&J, AstraZeneca, etc.

- This phenomenon is a natural consequence of …
  1. pivotal trials being typically powered at ≥80% (**steep** part of the power curve)
  2. **uncertainty** in the true treatment effect

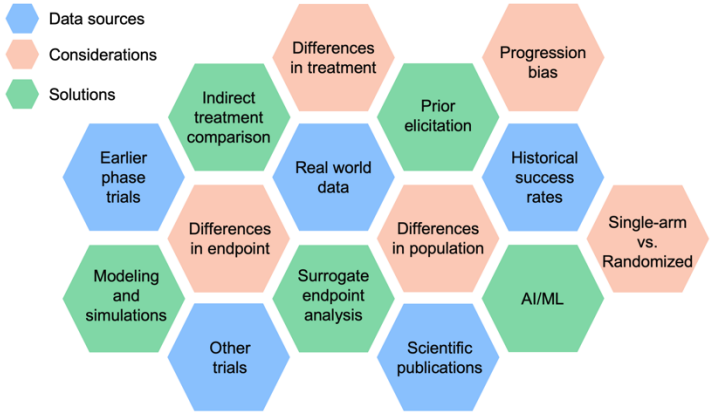


Takebe et al., (2018).

Typically, DDCP < Power

Oftentimes, prior knowledge needs to be translated to information on the Ph3 treatment effect.

This introduces **additional uncertainty!**



Typically, DDCP < Power

Instead of achieving statistical significance (i.e., meeting MDD), we may consider other definitions of “success”.

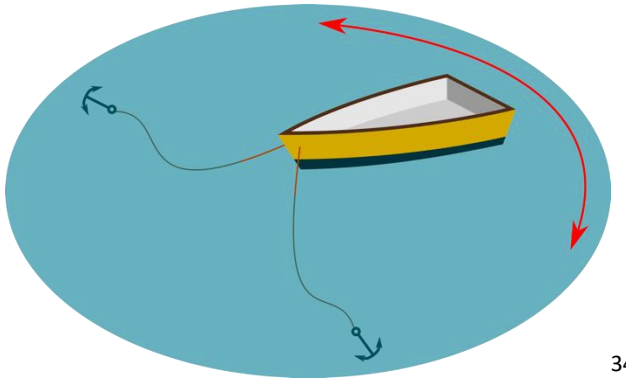
Example: observing an effect that is even better, such as meeting TPP

Higher bar to clear means **lower DDCP!**

“Data driven conditional probability (DDCP) is the probability of observing an estimated treatment effect that is **statistically significant**… given what we know about the true treatment effect.”

Define “success” clearly!

- Statistical significance? Clinical meaningfulness? MVP? TPP?
  - Typically recommend aligning first three and calculating **DDCP relative to MVP *and* TPP**
- Primary endpoint or more?
  - PTS = prob. of progressing from one decision point to the next[1]
  - cPTS = clinical PTS, prob. of meeting efficacy *and* safety
    - **DDCP** considers Ph3 primary efficacy target and may serve as an **“anchor”** for cPTS[2]
    - Other considerations: safety, operations, etc.[3]
  - rPTS = regulatory PTS, prob. of gaining approval
  - POL = overall prob. of launch



[1] PTS Handbook
[2] Statistics Boost 9
[3] Project Providentia

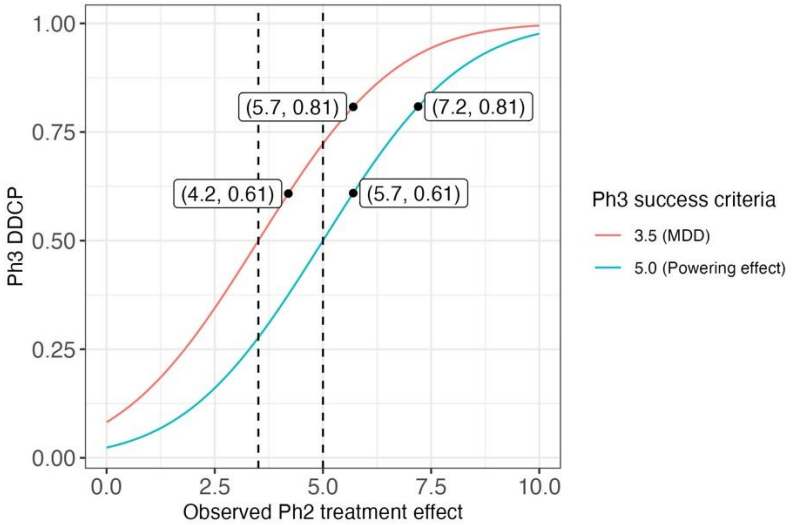
GYM329 – Potential Ph3 DDCP

Proposed Ph2 trial design:

- Percent body weight loss at 1y
- 1:1:1:1 randomization, 52 patients/arm

Question: What would we need to see in Ph2 for Ph3 to have 60% DDCP?

Note: Actual ”success” will require meeting other efficacy and safety endpoints, so trial will be much larger

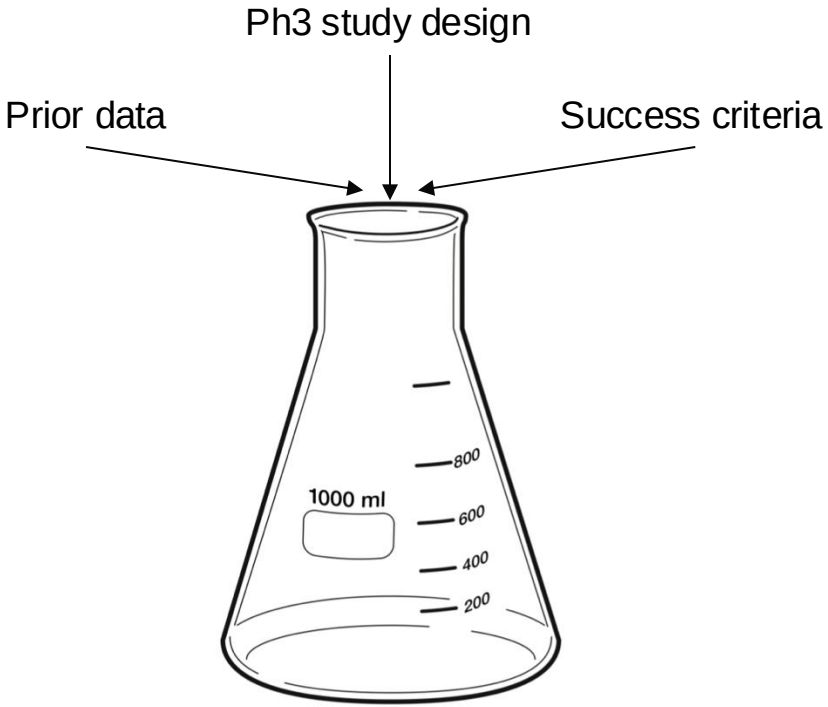


Today's program

"If our Ph3 studies are powered at 80%, why aren't we seeing 80% success rates?"

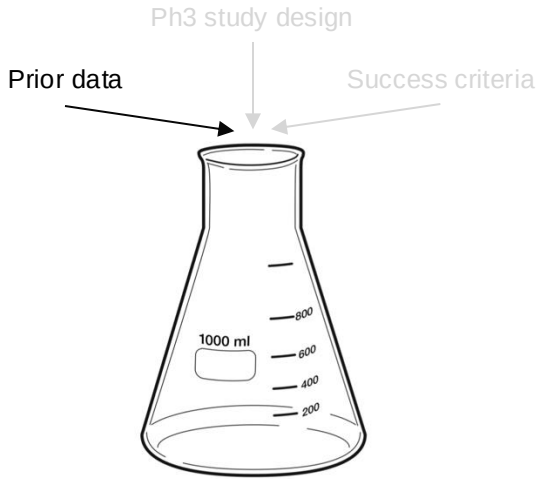
- What does it mean to power a study at 80%?
- If a study is powered at 80%, what are the chances of it turning out positive?
- How can we increase Ph3 PTS?

Ingredients of DDCP



How can we increase PTS?

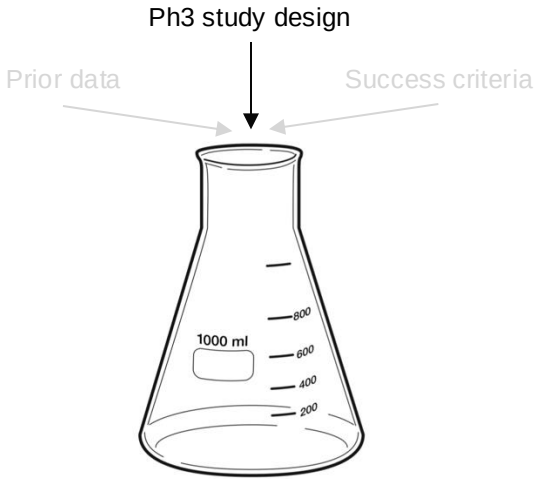
- Improve **prior data** by generating a robust and sufficiently large early data package
  - Tradeoff between …
    - robustness vs. resources/time
    - learning from early phase data vs. minimizing differences between phases
  - Each phase should be designed to generate data that are actionable for the next phase
  - If other sources of data are used, then …
    - minimize bias
    - be transparent with assumptions
    - understand risk and reflect uncertainty



How can we increase PTS?

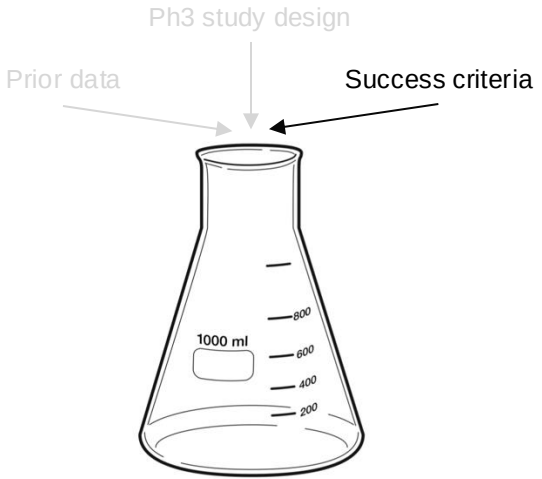
Improve **Ph3 study design** by increasing sample size, thus increasing power (e.g., 80% to 90%).

- Idea: DDCP < 90%
- However, no guarantee that DDCP will increase
- Also, increasing sample size relaxes the MDD and risks an outcome that is significant but not clinically meaningful (i.e., “so what?” scenario)



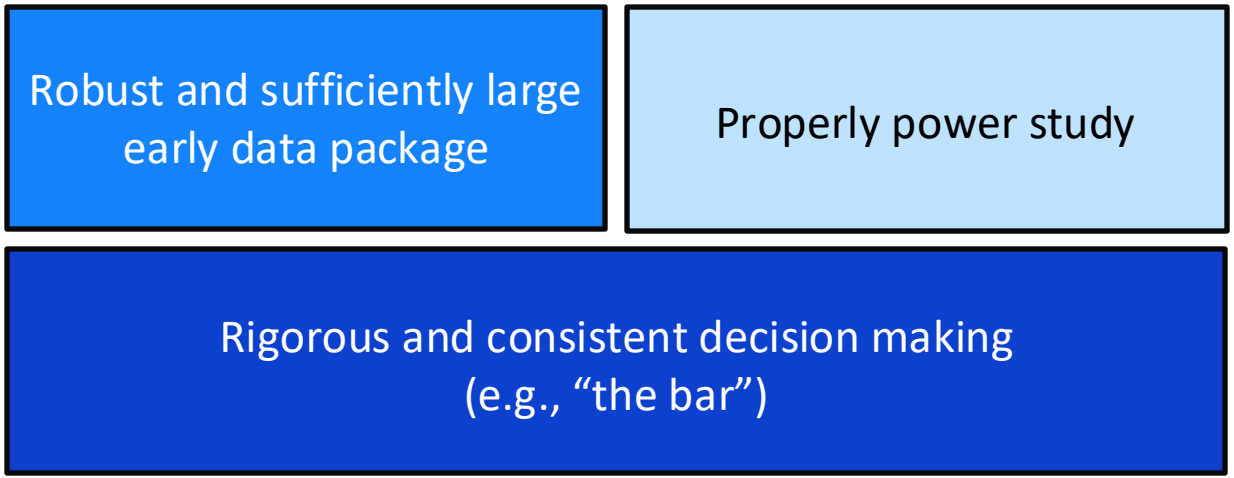
How can we increase PTS?

- Lower Ph3 **success criteria**.
  - Idea: Increase PTS by making ”success” easier to achieve
  - Not based on clinical relevance or commercial value
  - Not in line with R&D aspirations for high impact medicines



How can we increase Ph3 PTS?

3 key measures



Takeaways

“If our Ph3 studies are powered at 80%, why aren’t we seeing 80% success rates?”

- By acknowledging and incorporating **uncertainty about the true treatment effect**, DDCP illustrates why a Ph3 trial’s probability of “success” (“success” defined as statistical significance) is typically less than the trial’s statistical power.

- **Define “success” clearly!**

- The most important measures we can take to increase PTS are **generating robust and sufficiently large early data packages** and **applying a consistent decision framework**.

Appendix

Minimal detectable difference (MDD)

- "Smallest observed treatment effect for which trial will be significant"
- Ideally, we want to pick Ph3 sample size such that

"statistical significance" ⟺ "clinical relevance"

- Case 1:

"So what?": Statistically significant, but not clinically relevant

- Case 2:

"Missed opportunity": Not statistically significant, but clinically relevant

- Case 3:

"Always aligned": MDD aligned with minimal TPP

Doing now what patients need next