Statistics Boost Advanced 8

Data driven conditional probability: what and how?

Godwin Yung, PhD

Thanks:

Other members of the Methods, Collaboration and Outreach group (MCO)

Ulrich Beyer

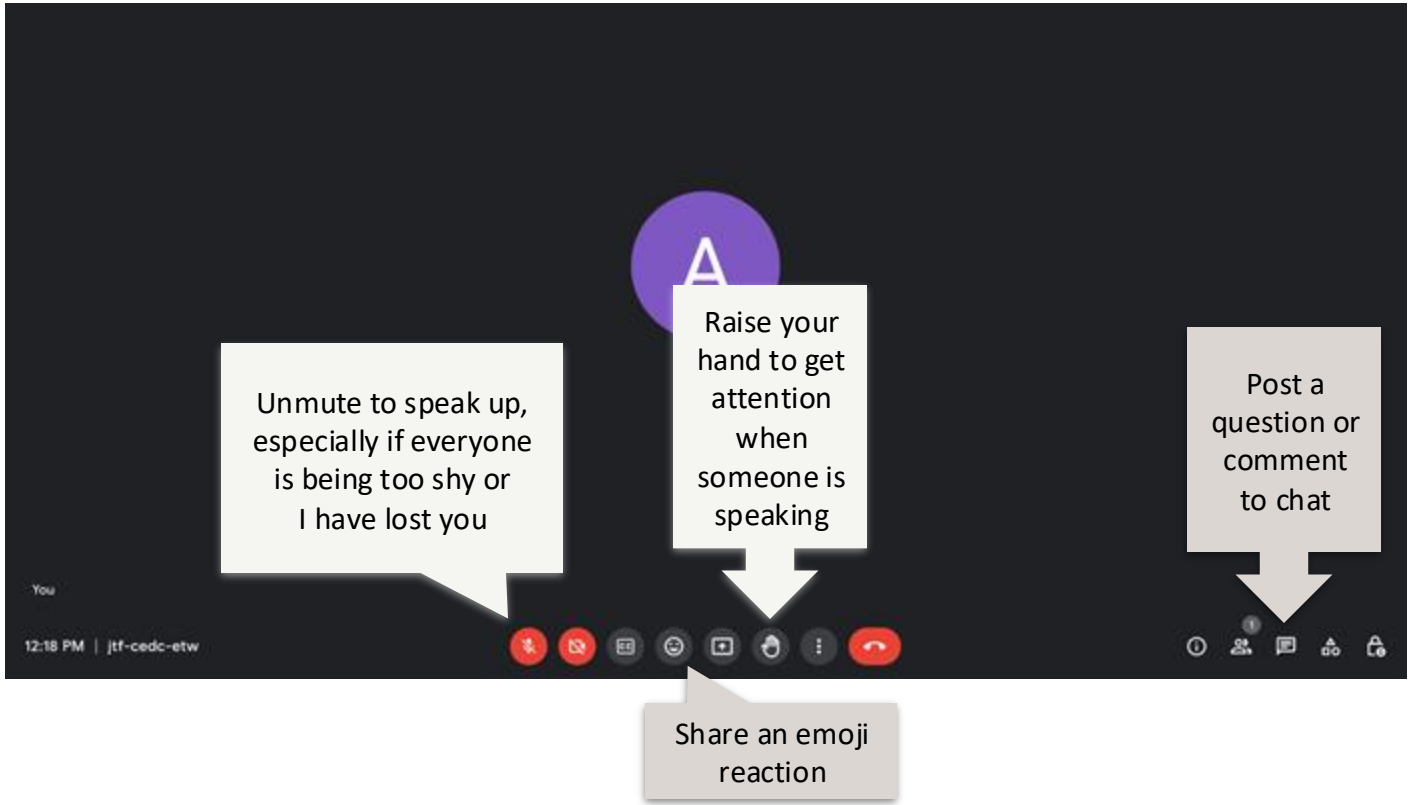
Statistics Boost series – since 2022

- Aim to upskill **the entire organization** in a non-technical way on relevant statistics topics and how to apply them effectively in drug development.
- **Non-technical** interactive sessions, 1h/month
- Topics & subscription information: PD connect post
- Recordings & slides of past sessions: Nucleus video channel
- Next session: Confirmatory Adaptive Designs (April 28th)

Statistics Boost Advanced series – since 2024

- Aim to upskill **statisticians** (and other colleagues with quantitative background) on relevant statistical methodology
- **Technical** interactive sessions, 1h/month
- Topics & subscription information: PD Connect post
- Recordings & slides of past sessions: Nucleus video channel
- Next session: Continuous longitudinal endpoints in clinical trials (May 12th)

Interacting with each other



Today's program

- **What** *exactly* is data driven conditional probability?
- **How** to specify the prior distribution?
  1. Ph2 trial
  2. Other trials
  3. Historical benchmark
  4. Interim analysis
- DDCP and PTS at **Roche**

Today's program

- **What** *exactly* is data driven conditional probability?
- **How** to specify the prior distribution?
  1. Ph2 trial
  2. Other trials
  3. Historical benchmark
  4. Interim analysis
- DDCP and PTS at **Roche**

What is data driven conditional probability (DDCP)?

"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"

Source: Statistics Boost 21

What is data driven conditional probability (DDCP)?

It is a probabilistic statement

"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"

Source: Statistics Boost 21

What is data driven conditional probability (DDCP)?

Ingredient #1: a success criteria

"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"

Source: Statistics Boost 21

What is data driven conditional probability (DDCP)?

Ingredient #2: prior knowledge of the treatment effect

"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"

Source: Statistics Boost 21

What is data driven conditional probability (DDCP)?

Ingredient #3: a Ph3 trial design

"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"

Source: Statistics Boost 21

What *exactly* is DDCP?

- θ = treatment effect of interest (i.e., the estimand)
- $\hat{\theta}$ = estimate of θ from a Ph3 trial
  - E.g., $\hat{\theta} \sim N(\theta, 4\sigma^2/n)$ for difference in means, assuming equal variance
- θ_{suc} = success criteria for $\hat{\theta}$ (e.g., TPP, MVP, statistical significance)
- $\pi(\cdot)$ = distribution of θ (i.e., the "prior", which is informed by existing data)

$$\text{DDCP} = E_\theta[\Pr_\theta(\hat{\theta} > \theta_{suc})]$$

Discrete θ Continuous θ

$$\sum_{i=1}^{n} \Pr_{\theta=\theta_i}(\hat{\theta} > \theta_{suc}) \times \pi(\theta_i) \quad \text{or} \quad \int \Pr_\theta(\hat{\theta} > \theta_{suc}) \times \pi(\theta)\, d\theta$$

GYM329 in obesity

Proposed Ph3 trial design[1]

- Estimand:
  - Population: adult patients with obesity
  - Treatment: GYM329 + incretin vs. incretin
  - Endpoint/Summary measure: Difference in percent body weight loss at 1y
- Significance level: one-sided 2.5%
- Power: 90% to detect a 5.0% difference
- Sample size: 69 patients/arm[2]
- Minimal detectable difference (MDD): 3.0% difference


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

GYM329 in obesity

Proposed Ph3 trial design[1]

- Estimand:
  - Population: adult patients with obesity
  - Treatment: GYM329 + incretin vs. incretin
  - Endpoint/Summary measure: Difference in percent body weight loss at 1y
- Significance level: one-sided 2.5%
- Power: 90% to detect a 5.0% difference
- Sample size: 138 patients, randomized 1-to-1
- Minimal detectable difference (MDD): 3.0% difference

θ (what $\hat{\theta}$ is trying to estimate)

GYM329 in obesity

Proposed Ph3 trial design[1]

- Estimand:
  - Population: adult patients with obesity
  - Treatment: GYM329 + incretin vs. incretin
  - Endpoint/Summary measure: Difference in percent body weight loss at 1y
- Significance level: one-sided 2.5%
- Power: 90% to detect a 5.0% difference
- Sample size: 138 patients, randomized 1-to-1
- Minimal detectable difference (MDD): 3.0% difference

Desired false positive and true positive rates

GYM329 in obesity

Proposed Ph3 trial design[1]

- Estimand:
  - Population: adult patients with obesity
  - Treatment: GYM329 + incretin vs. incretin
  - Endpoint/Summary measure: Difference in percent body weight loss at 1y
- Significance level: one-sided 2.5%
- Power: 90% to detect a 5.0% difference
- Sample size: 138 patients, randomized 1-to-1
- Minimal detectable difference (MDD): 3.0% difference

Standard error of $\hat{\theta}$ due to endpoint variability σ and finite sample size n

GYM329 in obesity

Proposed Ph3 trial design[1]

- Estimand:
  - Population: adult patients with obesity
  - Treatment: GYM329 + incretin vs. incretin
  - Endpoint/Summary measure: Difference in percent body weight loss at 1y
- Significance level: one-sided 2.5%
- Power: 90% to detect a 5.0% difference
- Sample size: 138 patients, randomized 1-to-1
- Minimal detectable difference (MDD): 3.0% difference

θ_{suc}

GYM329 in obesity

Proposed Ph3 trial design[1]

- Estimand:
  - Population: adult patients with obesity
  - Treatment: GYM329 + incretin vs. incretin
  - Endpoint/Summary measure: Difference in percent body weight loss at 1y
- Significance level: one-sided 2.5%
- Power: 90% to detect a 5.0% difference
- Sample size: 138 patients, randomized 1-to-1
- Minimal detectable difference (MDD): 3.0% difference

θ_{suc}

GYM329 in obesity

Ph2 trial (discrete case)

- Estimand: Same as Ph3
- Sample size: 45 patients/arm
- Outcome: 3.5% difference

Treatment effect	Probability of being true
1.0%	0.25
3.5%	0.50
6.0%	0.25

GYM329 in obesity

Ph2 trial (discrete case)

$\pi(\cdot)$ $\quad$ $\Pr_\theta(\hat{\theta} > \theta_{suc})$

Treatment effect	Probability of being true	Probability of Ph3 success if true
1.0%	0.25	10%
3.5%	0.50	63%
6.0%	0.25	97%

$$\text{DDCP} = \sum_{i=1}^{3} \Pr_{\theta=\theta_i}(\hat{\theta} > \theta_{suc}) \times \pi(\theta_i)$$

GYM329 in obesity

Ph2 trial (discrete case)

$\pi(\cdot)$ $\quad$ $\Pr_\theta(\hat{\theta} > \theta_{suc})$

Treatment effect	Probability of being true	Probability of Ph3 success if true
1.0%	0.25	10%
3.5%	0.50	63%
6.0%	0.25	97%

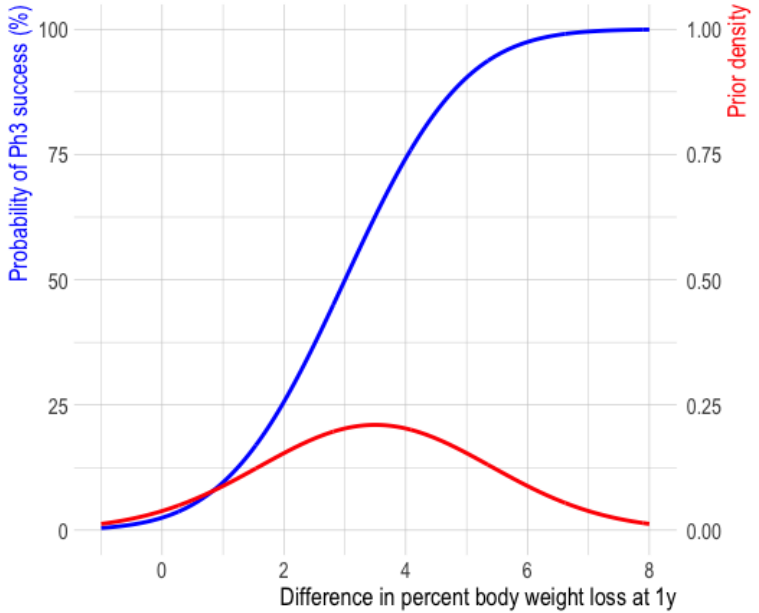
DDCP = 10% x 0.25 + 63% x 0.50 + 97% x 0.25 = 58%

GYM329 in obesity

Ph2 trial (continuous case)

- Estimand: Same as Ph3
- Sample size: 49 patients/arm
- Outcome: 3.5% difference
  - 80% CI (1.1, 5.9)

$$\text{DDCP} = \int \Pr_{\theta}\left(\hat{\theta} > \theta_{suc}\right) \times \pi(\theta)\, d\theta$$
$$= 58\%$$



R code

Using bpp for basic DDCP calculations

SE from Ph2 →

```
> library(bpp)
> mdd       <- 3
> stDev     <- 9
> n_Ph3     <- 69   # per arm
> n_Ph2     <- 49   # per arm
> theta_Ph2 <- 3.5
> sigma_Ph2 <- stDev / sqrt(n_Ph2 / 2)
> bpp_continuous(prior = "normal",
+                successmean = mdd,
+                stDev = stDev,
+                n1 = n_Ph3, n2 = n_Ph3,
+                priormean = theta_Ph2,
+                priorsigma = sigma_Ph2)
[1] 0.5832747
```

← Success criteria

← Ph3 design

← Prior distribution

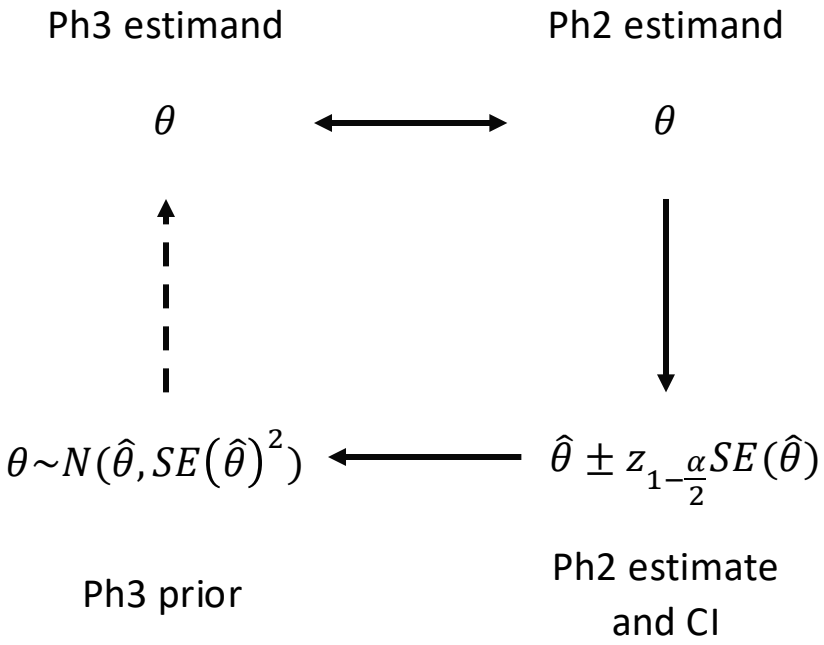
Today's program

- **What** *exactly* is data driven conditional probability?
- **How** to specify the prior distribution?
  1. Ph2 trial
  2. Other trials
  3. Historical benchmark
  4. Interim analysis
- DDCP and PTS at **Roche**

1. An ideal Ph2 trial, the “small Ph3”

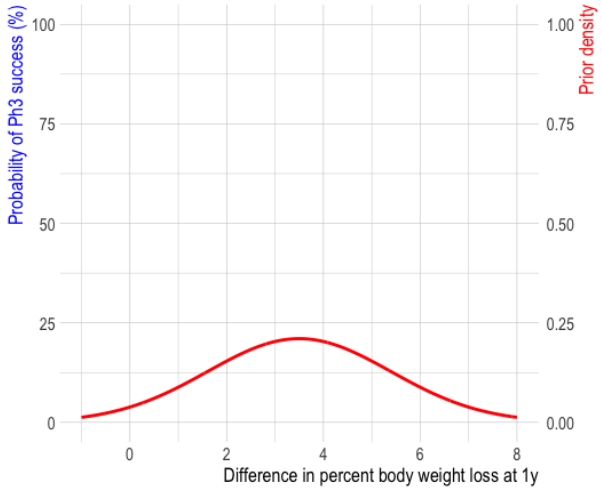
Ph2 trial estimand = Ph3 trial estimand

- Population
- Treatments
- Endpoint/Summary measure
- (Intercurrent events)



GYM329 in obesity

- Ph2 outcome: 3.5% difference (80% CI [1.1, 5.9])
  - $\hat{\theta} = 3.5$
  - $\alpha = 0.2$ (two-sided)
  - $z_{1-\frac{\alpha}{2}} = 1.28$
  - $SE(\hat{\theta}) = 1.9$
- Ph3 prior: $\theta \sim N(3.5, 1.9^2)$



2. Other trials

- Existing trial estimand ≠ Ph3 trial estimand

Attribute	Considerations	Potential solutions
Population	If there are differences in population, do we expect treatment benefit to be the same between the two populations? Was the motivation for the Ph3 population based on a subgroup finding?	matching adjusted indirect comparisons (MAIC), propensity score matching (PSM) multiplicity adjustment, shrinkage estimation
Treatment	Randomized with same comparator? Single-arm? Time-trend for standard of care? Do we expect similar benefit for treatments with similar target and mechanism of action (MoA)? What about treatments with different target and/or MoA?	Bucher method, network meta-analysis, statistical models
Endpoint/Summary measure	If there are differences in endpoint, do we know how the different endpoints are related?	meta-analysis, surrogacy, disease area prediction models (e.g., TGI-OS, multistate models, longitudinal predictions), simulations

2. Other trials

- Existing trial estimand ≠ Ph3 trial estimand

Attribute	Considerations	Potential solutions
Population	If there are differences in population, do we expect treatment	matching adjusted indirect s (MAIC), propensity hing (PSM)
		adjustment, shrinkage
Treatment		thod, network meta- atistical models
	Do we expect similar benefit for treatments with similar target and mechanism of action (MoA)? What about treatments with different target and/or MoA?	
Endpoint/Summary measure	If there are differences in endpoint, do we know how the different endpoints are related?	meta-analysis, disease area prediction models (e.g., multi-state models, surrogacy, longitudinal predictions), simulations

Disclaimer:
Since the prior distribution (mean and variance) is a critical input of DDCP, careful assessment of the plausibility of assumptions underlying these non-standard derivations of a prior is required.

3. Historical benchmark

- Historical benchmarks ask, “How have other trials in a *similar* situation fared?”
  - Study phase (e.g., I, II, III, III after skipping II)
  - Therapeutic area (e.g., cardiovascular, oncology)
  - Life cycle (e.g., new molecular entity, line extension)

- **Why** is it important to incorporate historical benchmarks into the prior?

Question #1

Ph2 outcome: **5% difference**

What will DDCP trend towards as Ph2 sample size decreases?

a. 0%
b. 50%
c. 100%

Ph2 sample size (per arm)	DDCP
49	80%
30	?
20	?
10	?
1	?

Question #2

Ph2 outcome: **0% difference**

What will DDCP trend towards as Ph2 sample size decreases?

a. 0%
b. 50%
c. 100%

Ph2 sample size (per arm)	DDCP
49	10%
30	?
20	?
10	?
1	?

Question #1

Ph2 outcome: **5% difference**

What will DDCP trend towards as Ph2 sample size decreases?

a. 0%

b. **50%**

c. 100%

Ph2 sample size (per arm)	DDCP
49	80%
30	76%
20	73%
10	68%
1	56%

Question #2

Ph2 outcome: **0% difference**

What will DDCP trend towards as Ph2 sample size decreases?

a. 0%

b. **50%**

c. 100%

Ph2 sample size (per arm)	DDCP
49	10%
30	14%
20	18%
10	24%
1	41%

Why is this happening?

- If all effect sizes along the real line are equally likely, then ~50% of effect sizes are above/below any success threshold.
  - What DDCP should be when there is little prior data should be context specific!
- Also concerning is the “large swings” in PTS despite little prior data.
- **”DDCP is not Bayesian”**: flat priors are still informative

Ph2 outcome based on 1 patient/arm	DDCP
-5.0	27%
-3.5	31%
0	41%
3.5	52%
5.0	56%

What to do about it?

1. Define a historical benchmark, i.e., ask “What’s our starting DDCP/PTS?”
   - Industry benchmarks (KMR)
   - Reference assumptions (2017 PTS Handbook)
   - Providentia ”Development Approach” risk factor

   Note: It’s important to keep in mind the success criteria of the historical benchmark!

2. Decide on an appropriate effective sample size (ESS) to be associated with the historical benchmark, i.e., ask “How informative is the starting DDCP/PTS relative to trial data?”
   - ESS = 10
   - ESS = 5% of Ph3 sample size

3. Back-engineer prior distribution so that DDCP (using only this prior as available information) equals the historical benchmark defined in Step 1.
   - Normal conjugate priors for normally distributed endpoints

R code

Adding an initial prior based on historical benchmark

```
> library(bpp)
> mdd       <- 3
> stDev     <- 9
> n_Ph3     <- 69   # per arm
> n_Ph2     <- 49   # per arm
> theta_Ph2 <- 3.5
> sigma_Ph2 <- stDev / sqrt(n_Ph2 / 2)
> bpp_continuous(prior = "normal",
+                successmean = mdd,
+                stDev = stDev,
+                n1 = n_Ph3, n2 = n_Ph3,
+                priormean = theta_Ph2,
+                priorsigma = sigma_Ph2)
[1] 0.5832747
```

R code

A. Back-calculate mean of initial prior

```
> library(bpp)
> mdd       <- 3
> stDev     <- 9
> n_Ph3     <- 69   # per arm
```

```
> benchmark     <- 0.6  # Step 1
> n_bench       <- 10   # Step 2
> sigma_bench   <- stDev / sqrt(n_bench) # Step 3
> bpp_continuous(prior = "normal",
+                successmean = mdd,
+                stDev = stDev,
+                n1 = n_Ph3, n2 = n_Ph3,
+                priormean = 3.82,
+                priorsigma = sigma_bench)
[1] 0.6001319
```

Back-calculating the prior mean can be done via trial-and-error (as showcased above) or analytically (see Appendix)

R code

B. Update initial prior with Ph2 data

```
> library(bpp)
> mdd        <- 3
> stDev      <- 9
> n_Ph3      <- 69   # per arm
> benchmark   <- 0.6  # Step 1
> n_bench     <- 10   # Step 2
> sigma_bench  <- stDev / sqrt(n_bench) # Step 3
> bpp_continuous(prior = "normal",
+                successmean = mdd,
+                stDev = stDev,
+                n1 = n_Ph3, n2 = n_Ph3,
+                priormean = 3.82,
+                priorsigma = sigma_bench)
[1] 0.6001319
```

```
> theta_bench <- 3.82
> n_Ph2       <- 49   # per arm
> theta_Ph2   <- 3.5
> sigma_Ph2   <- stDev / sqrt(n_Ph2 / 2)
> prior_bench_Ph2 <- NormalNormalPosterior(
+   datamean = theta_bench,
+   sigma = sigma_bench,
+   n = 1,
+   nu = theta_Ph2,
+   tau = sigma_Ph2)
```

R code

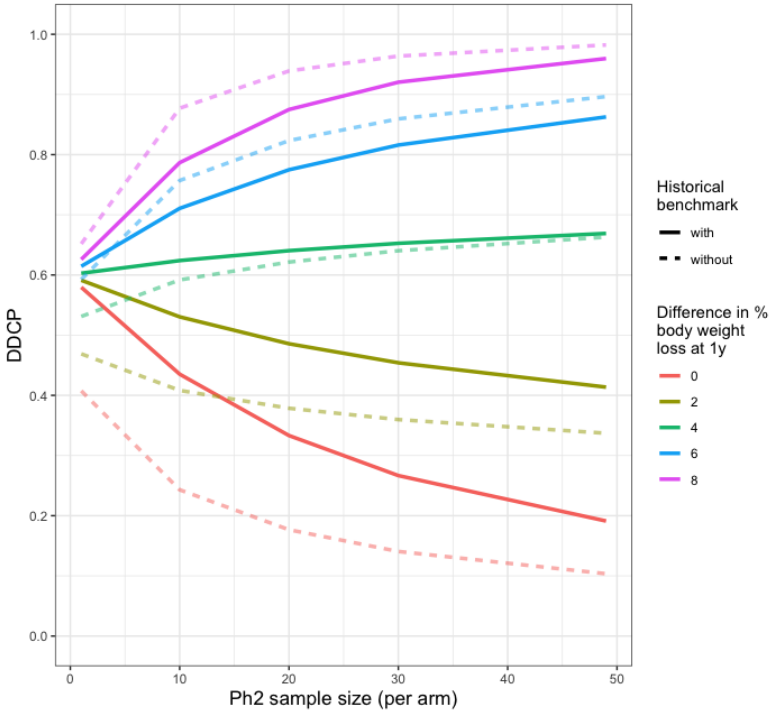
C. Calculate DDCP using updated prior

```
> library(bpp)
> mdd        <- 3
> stDev      <- 9
> n_Ph3      <- 69   # per arm
> benchmark   <- 0.6  # Step 1
> n_bench     <- 10   # Step 2
> sigma_bench  <- stDev / sqrt(n_bench) # Step 3
> bpp_continuous(prior = "normal",
+                successmean = mdd,
+                stDev = stDev,
+                n1 = n_Ph3, n2 = n_Ph3,
+                priormean = 3.82,
+                priorsigma = sigma_bench)
[1] 0.6001319
> theta_bench <- 3.82
> n_Ph2       <- 49   # per arm
> theta_Ph2   <- 3.5
> sigma_Ph2   <- stDev / sqrt(n_Ph2 / 2)
> prior_bench_Ph2 <- NormalNormalPosterior(
+   datamean = theta_bench,
+   sigma = sigma_bench,
+   n = 1,
+   nu = theta_Ph2,
+   tau = sigma_Ph2)
```

```
> bpp_continuous(prior = "normal",
+                successmean = mdd,
+                stDev = stDev,
+                n1 = n_Ph3, n2 = n_Ph3,
+                priormean = prior_bench_Ph2$postmean,
+                priorsigma = prior_bench_Ph2$postsigma)
[1] 0.6077822
```

Why is it important to incorporate historical benchmarks into the prior?

- Provide a meaningful starting DDCP/PTS
- Make clear when we have limited prior data
- Protect against ”winner’s curse” (tendency for successful early phase trials to overestimate the true effect size)



4. Interim analysis

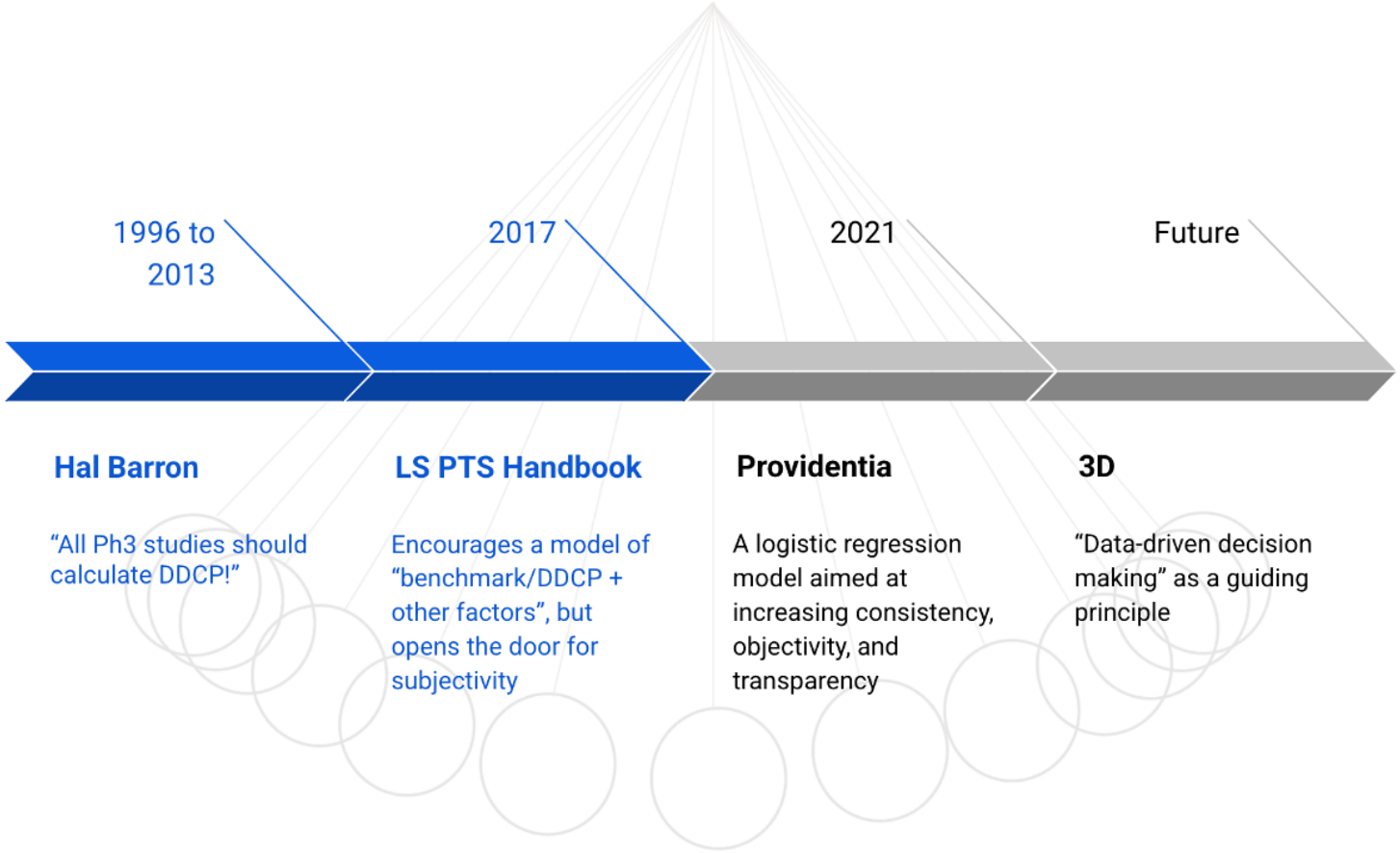
- Technically, an efficacy or futility interim analysis is an **ideal situation to update the prior**.
  - Ph3 estimand at interim = Ph3 estimand at final!
  - Tutorials available on successR

- Whether or not to update the prior and DDCP following an interim analysis is more of a **philosophical debate**.
  - \+ Better manage expectations and resources
  - \- Trial will continue anyway
  - \- Passing a non-binding futility analysis does not guarantee passing the futility criteria
  - \- One might be able to back-calculate the futility boundary given change in DDCP

Today's program

- **What** *exactly* is data driven conditional probability?
- **How** to specify the prior distribution?
  1. Ph2 trial
  2. Other trials
  3. Historical benchmark
  4. Interim analysis
- DDCP and PTS at **Roche**

To calculate or not to calculate?



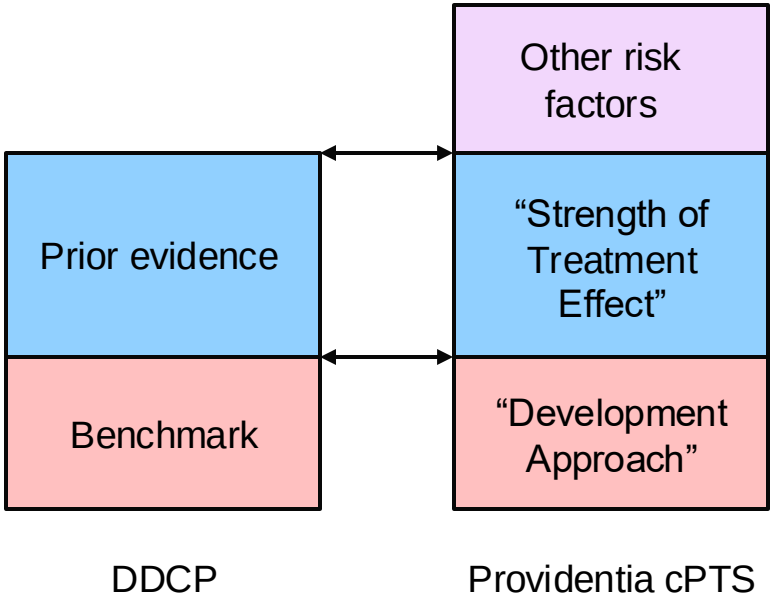
DDCP vs. cPTS

- Data driven conditional probability (DDCP): probability that a Ph3 study will pass a specific **efficacy threshold**
  - A critical components is the **prior distribution for the Ph3 treatment effect**

- Clinical probability of technical success (cPTS): probability of a Ph3 study achieving **clinical success, thereby moving on to filing**
  - Considerations include not only prior evidence of clinical efficacy, but also safety risks, operational risks, prior evidence supporting target dose and schedule, etc.
  - **Providentia** is a logistic regression model that estimates cPTS based on **15 qualitative risk factors**

Providentia v2.0 (coming 2H2025)

Providentia cPTS ≈ DDCP + other risk factors



Takeaways

Takeaways

1. DDCP is relevant and should be calculated when possible, even in addition to other measures of cPTS

2. DDCP is sensitive to inputs, so think carefully about this in relation to the information available

3. Even though calculating DDCP is a tall task, statisticians are well-positioned to drive better quantitative decision making

Learning resources

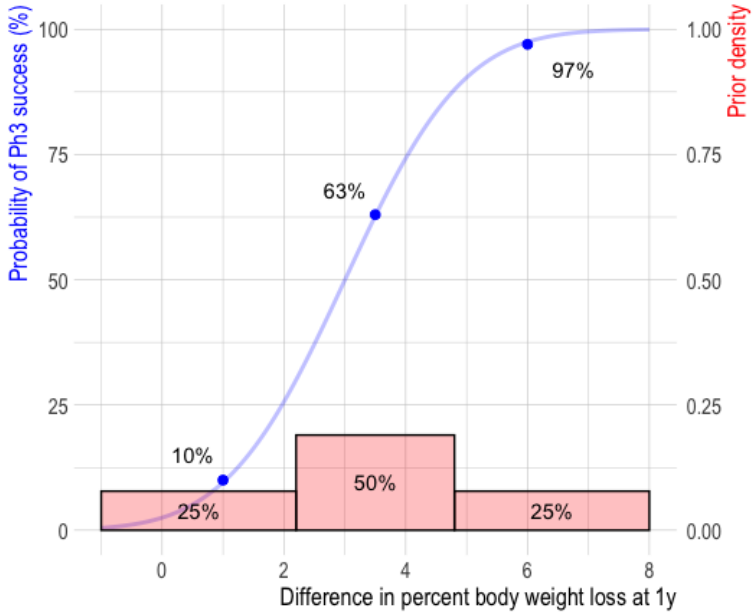
- successR
  - Training material on DDCP
  - Tutorials (back-engineering a prior, updating DDCP after interim, other prior choices, etc.)
- 2017 Late Stage PTS Handbook
- Providentia
- Rufibach, K. et al. (2016) Bayesian predictive power: choice of prior and some recommendations for its use as probability of success in drug development. *Pharm Stat 15*(5):438-446.
- Hampsion, L. et al. (2021) Improving the assessment of the probability of success in late stage drug development. *Pharm Stat.* 21:439-459.

Appendix

Using Riemann sums to approximate DDCP

- Sometimes, in practice, simulations are required to calculate probability of success (PoS) given a fixed treatment effect. In such cases, integrating probability of success over a continuous prior distribution (i.e., large number of treatment effects) may be computationally intensive.

- One could instead use a Riemann sum to approximate DDCP.
  1. Partition prior into discrete prob's of potential treatment effects.
  2. Assign a PoS to each partition.



Normal conjugate priors

If the treatment effect estimate and prior both follow normal distributions, that is, $\hat{\theta} \sim N\left(\theta, \sigma_{fin}^2\right)$ and $\theta \sim N(\theta_0, \sigma_0^2)$, then DDCP can be written explicitly as

$$DDCP = E_\theta\left[\Pr_\theta\left(\hat{\theta} > \theta_{suc}\right)\right] = \Phi\left(\frac{\theta_{suc} - \theta_0}{\sqrt{\sigma_{fin}^2 + \sigma_0^2}}\right)$$

This equation can be used to back-calculate the prior mean θ_0 given a desired DDCP/cPTS. Indeed, by algebraic manipulation, we have that

$$\theta_0 = \theta_{suc} - \Phi^{-1}(DDCP)\sqrt{\sigma_{fin}^2 + \sigma_0^2}$$

Doing now what patients need next