# Conceivable design options for Ph2/Ph3 in heart failure

Axel Wetterlundh, Yuejia Xu, Alexander Bore, Marcus Millegård

# Agenda

- Introduction to the compared development options
- Literature review and applications on the domain framework
- Presentation of the Bayesian futility track, including comparison of including/excluding Phase 2B
- Conclusion

# Acknowledgments

We would like to thank our collaborators for highly valuable input from different AstraZeneca departments

Early Cardiovascular, Renal & Metabolism (CVRM) – Malin Aurell, Anders Gabrielsen

Clinical operations – Ann Nilsson

Late CVRM Biometrics –Samvel Gasparyan, Karin Gustafsson, Per Nyström

Patient Centered Science – Folke Folkvaljon

Early Biometrics & Statistical Innovations – Magnus Kjaer

Clinical Pharmacology and Quantitative Pharmacology – Ann-Charlotte Egnell, Bengt Hamrén, Magnus Åstrand

# Design of clinical programs in heart failure

# The importance of a well-designed clinical program

***"You have a great portfolio, now you need to bring medicines to patients"***

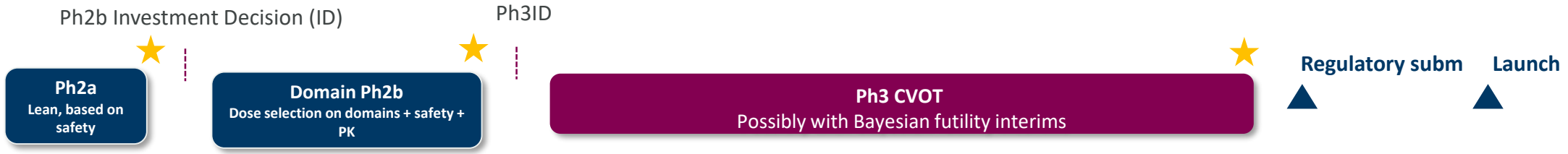
– AstraZeneca CEO Pascal's comment on early Cardiovascular, Renal & Metabolism (CVRM) portfolio during spring Scientific review

- "Bringing medicines to patients" ⇔ optimizing clinical development programs
  - Put patients first
  - **Maximize number of successful Phase 3 trials for a given time frame and budget**

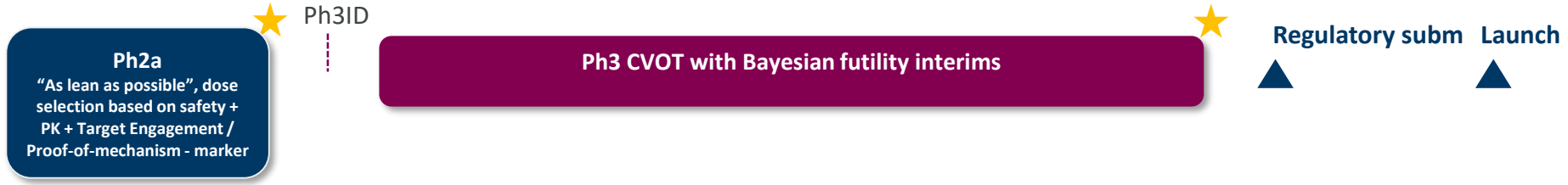
# 2 conceivable clinical development plans in Heart Failure

- The heart failure therapy area is lacking a validated and reliable Phase 2 surrogate endpoint that predicts Phase 3 results with high accuracy
- Potential design options:

**Option 1: Domain track** - use domain based approach to increase probability of successful Cardiovascular outcome trial (CVOT)




**Option 2: Bayesian futility track** - test efficacy in Ph3




- **Aim: A fair and realistic comparison of the options with respect to time and cost to maximize number of successful compounds and improve future development plans**

# Domain track

**Aim:** evaluate the use of domains in Ph2B as decision making for Ph3

# Domains and variables included in the domain approach

1. **Biomarker domain**
   - N-terminal pro B-type natriuretic peptide (**NT-proBNP**)
2. **Exercise capacity domain**
   - 6-minute walking distance (**6MWD**)
   - Maximal oxygen consumption (**VO2max**)
3. **Health-related quality of life domain**
   - Kansas City Cardiomyopathy Questionnaire – total symptom score (**KCCQ-TSS**)
4. **Cardiac structure and function (imaging) domain**
   - Global longitudinal strain (**GLS**)
   - Left atrial volume index (**LAVI**)
   - Left ventricular mass index (**LVMI**)
   - Left ventricular ejection fraction (**LVEF**)

Remark: we have also experimented with adding the “clinical events” domain (slightly increase false positives). Today’s presentation will focus on the case without the “clinical events” domain.

# Domain-based GNG framework & data collection

For each variable: We follow the decision framework described in Frewer et al. (2016)[1]

- Definition of target value (TV): desired level of effect
- Definition of lower reference value (LRV): minimal level of effect
- GO if: Probability of being worse than LRV < 20%
- STOP if: Probability of being better than TV < 10% (STOP overrides GO)

For each domain: GO for the domain if ≥ 1 variable(s) in this domain achieves GO

Overall decision criteria:

- GO if: ≥ 2 domains GO & none of the variables included is statistically significant in the wrong direction
- STOP if: No domain GO
- DISCUSS: otherwise (eg. ≥ 1 domain with GO but one variable is statistically significant in the wrong direction)

## How does this domain-based GNG framework perform on HF compounds?

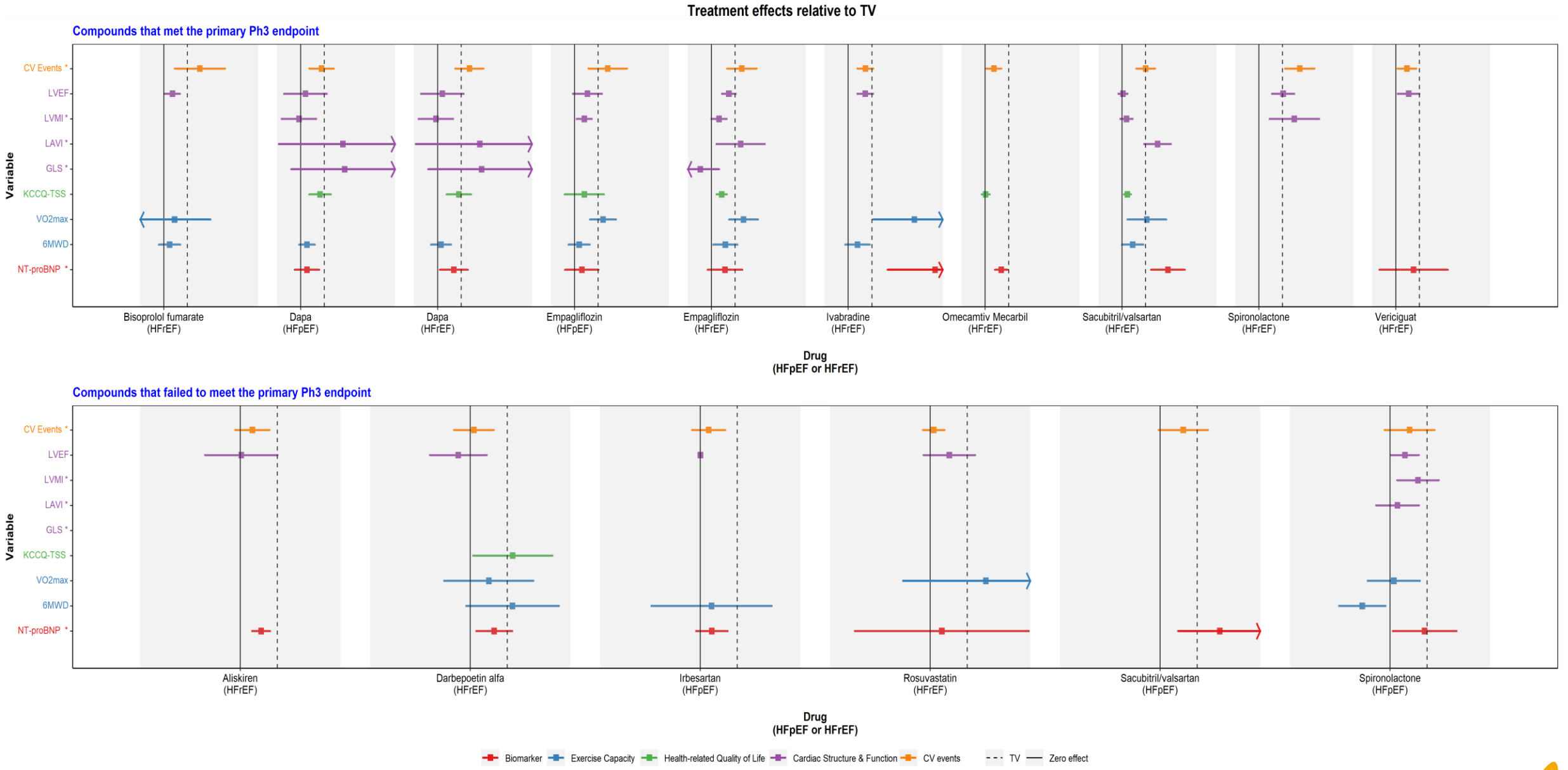
We have

1) Collected data on 12 compounds (by HFpEF and HFrEF, 19 in total) from 62 studies[2] within heart failure, with sufficient information on endpoints used in domains
2) Applied the domain decision framework to the data on these compounds



1. Frewer, P., Mitchell, P., Watkins, C., and Matcham, J. (2016) Decision-making in early clinical drug development. Pharmaceut. Statist., 15: 255– 263. doi: 10.1002/pst.1746.
2. Including Ph2, Ph3, and Ph4 studies

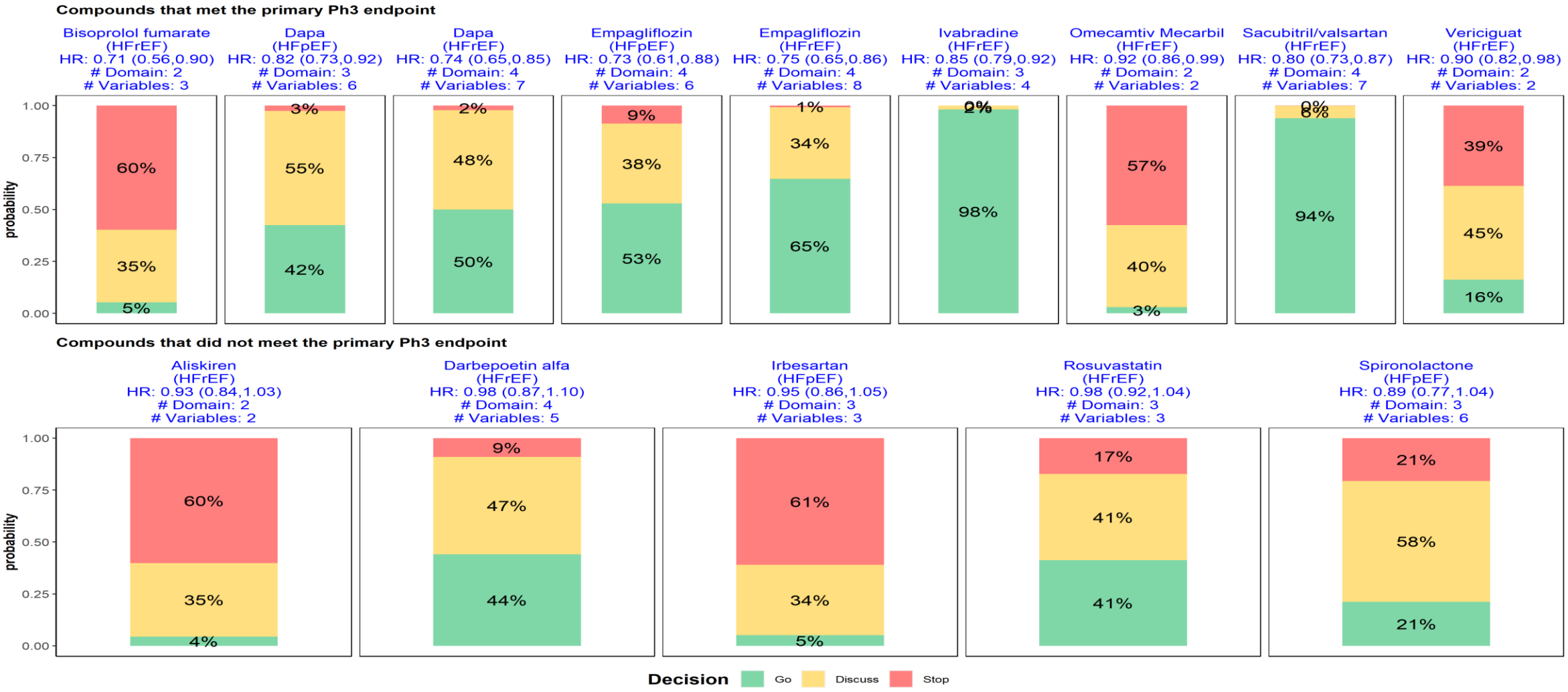
# Treatment effects (relative to TV) on domain variables




No clear pattern and no perfect predictor exists

# Observed effects scenario – simulation results

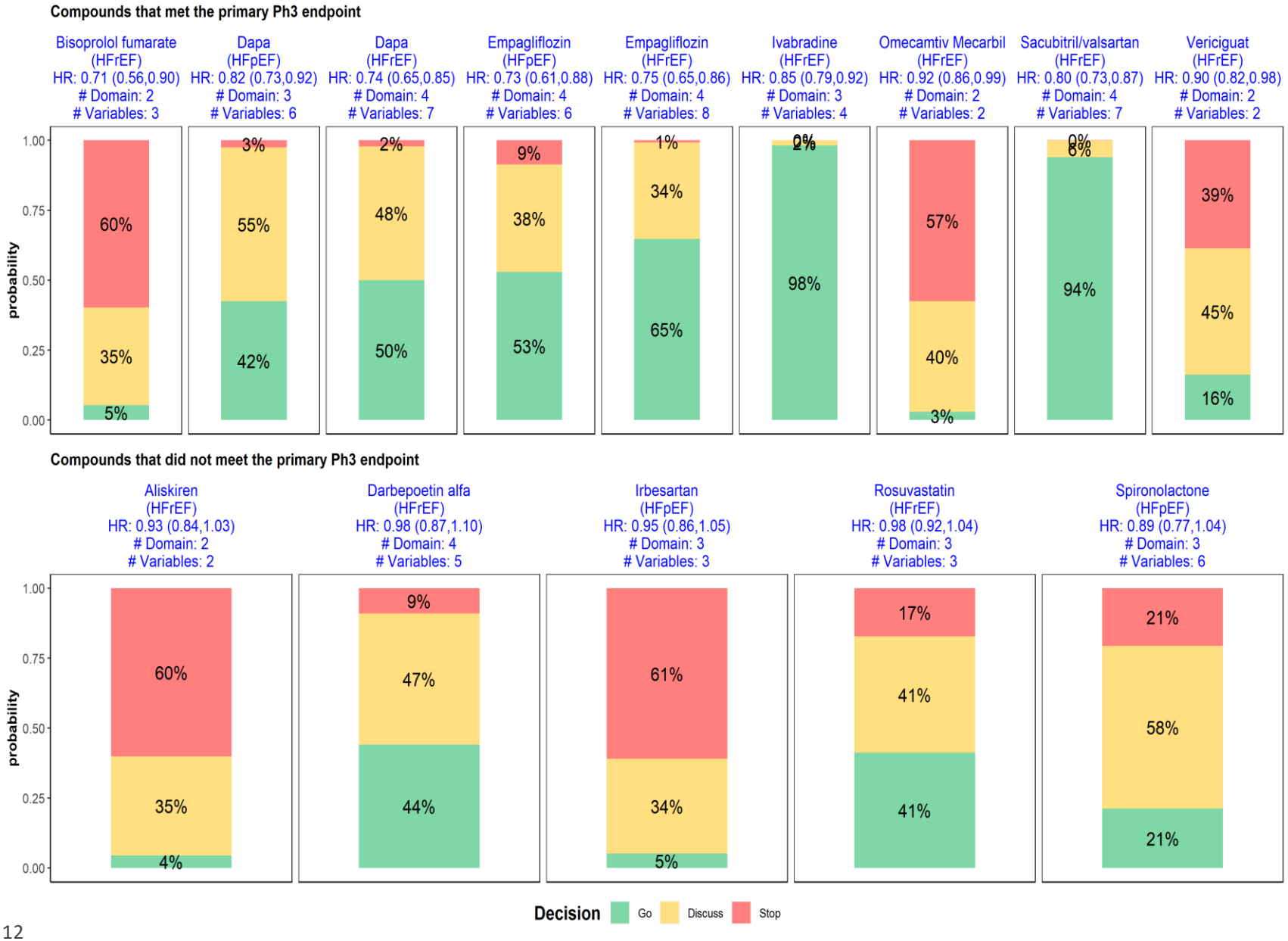
Given the treatment effects and variability (collected from the literature) on Ph2B endpoints for heart failure compounds, for each compound, we have simulated 1000 Ph2B trials, each with 150 patients/arm. We then apply the domain GNG decision framework to each simulated Ph2B trial, the probabilities of getting Go, Discuss , Stop among 1000 simulated trials are presented in the figure below:






**Sample size: 150 evaluable/arm. We have investigated other feasible Ph2B sample sizes, and patterns are similar.**

# Observed effects scenarios - interpretations




- Green bars are generally taller for compounds that met the primary Ph3 endpoints than those did not (except for BF, OM, and Vericiguat, see this)
- However, **the framework is not highly discriminative** (Dapa, Empa vs. Darbepoetin alfa and Rosuvastatin), and it could be tricky to "predict" Ph3 outcome based on the domain results from Ph2B
- Limitations of the analysis:
  - Only have information on a small number of compounds
  - Incomplete information on domain endpoints for some compounds
  - Not considering specific MoAs

# Domain track conclusion

- The domain framework is not highly discriminative between compounds that met the primary Ph3 endpoints and those did not

- Rough estimate (based on limited and incomplete data) of the correlation between predicted probability of GO by domain approach and Ph3 HR: 0.2~0.5

- We have investigated potential modifications of the domain framework
  - Not change the conclusion that “it is still challenging to make accurate GNG decisions based on a decision framework using multiple endpoints”
  - Some methods may help enhance power (by combining multiple endpoints) rather than construct an accurate GNG criteria



# Bayesian futility track

# The Bayesian approach in a nutshell

- **Aim:** to select an **optimal** interim futility **strategy** with a preserved or reduced type 1-error. The optimal strategy is the strategy which, within a given time frame and budget, yields the highest number of successful Ph3 trials

- At each interim, we calculate the **predictive probability**, that the one-sided p-value at the final analysis < 0.025, analyzed in the frequentist way. This probability is based on the data available at the interim and the prior distribution.

- If the predictive probability is smaller than a pre-specified threshold, then the study ends early for futility

- We have also examined whether it is worth to include a Ph2B
    - And if so, how predictive does the Ph2B need to be to be “worth the time and cost”?

# Description of a futility strategy

A strategy is built up by 3 components:

1. Number of interims
2. Time points for interims (based on the number of events)
3. Futility thresholds at the interims (based on predictive probabilities)

As an example, consider

1. # of interims: 3
2. Time point for interims: 100, 400, 600
3. Predictive probability thresholds at interims: 0.2, 0.1, 0.1

- Together the three components form a strategy

Even with some restrictions, a combination of the 3 components above yields ~100 000 different strategies

**How do we find the optimal strategy?**



# Goal: Maximize the number of studies that can be run within a certain time frame, with a certain budget

**In order to optimize, we first need to know what we are optimizing. We suggest basing it on the company's budget constraints**

If we had an infinite amount of money, we would run all studies with very low thresholds

- The profit of a successful compound by far outweighs the cost of the study, especially if it can be quick to launch
- Hence, we would run everything as quickly as possible
- In reality, we have a limited R&D budget and resource constraints, where spending time/money on a "lousy" compound may stop us from running a more promising one

Imagine we have a budget that can be translated to that we can only run one Ph3 study at the time (can just be scaled up)

Assume we have a bank of M number of simulated trials. Their hazard ratio effect is sampled from a distribution to reflect the company portfolio

Which one of all the ~100 000 strategies will be able to run the most successful studies within X days?

# Assumptions

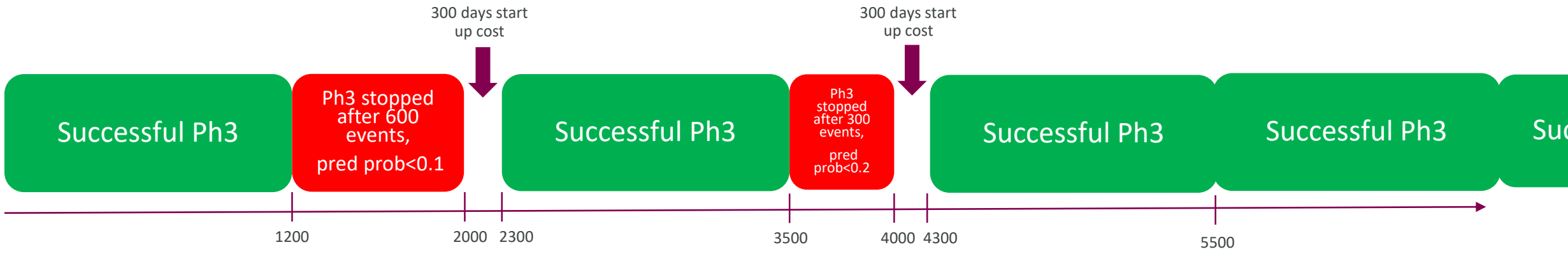
| Measure | Assumption | Explanation of assumption |
|---|---|---|
| Number of events in Ph3 | 1200 | |
| Successful trial | Statistically significant | Study successful if statistically significant ⇔<br>one-sided p-value < 0.025 ⇔<br>Hazard ratio estimate < 0.89 |
| Ph3 event rate, effect and recruitment pattern | Constant event rate and effect over time. Recruitment pattern follow a Carrol distribution (see picture in back-up) | |
| Ph3 time | 1200 days | |
| Ph3 cost | 200M USD | |
| Ph3 size | Event-driven, 1200 events, 2 arms (active vs control), 4000 patients | |
| Punishment for Ph3 early stopping | 300 days | To compensate for start-up time and costs of a Ph3 trial |
| Evaluated values of predictive probability thresholds | 0.1, 0.2, (0.3, 0.4) | |
| Evaluated number of interims | 0-7 | |
| Evaluated time points for interims | After 100, 200, 300, 400, 500, 600, 700, 800 or 900 events | |
| Evaluated “true” portfolio distributions of the hazard ratio | - Uniform between [0.7, 1]<br>- Diverse, uniform between [0.7, 0.75] with p=0.25, or uniform between [0.95, 1] with p=0.75 | |
| Evaluated prior distributions | ”Weak”: log(HR) ~N(0,1000) | |

See back-up for more assumptions specific for the comparison of including / not including Ph2B

# Illustration of the simulations

Days left:0

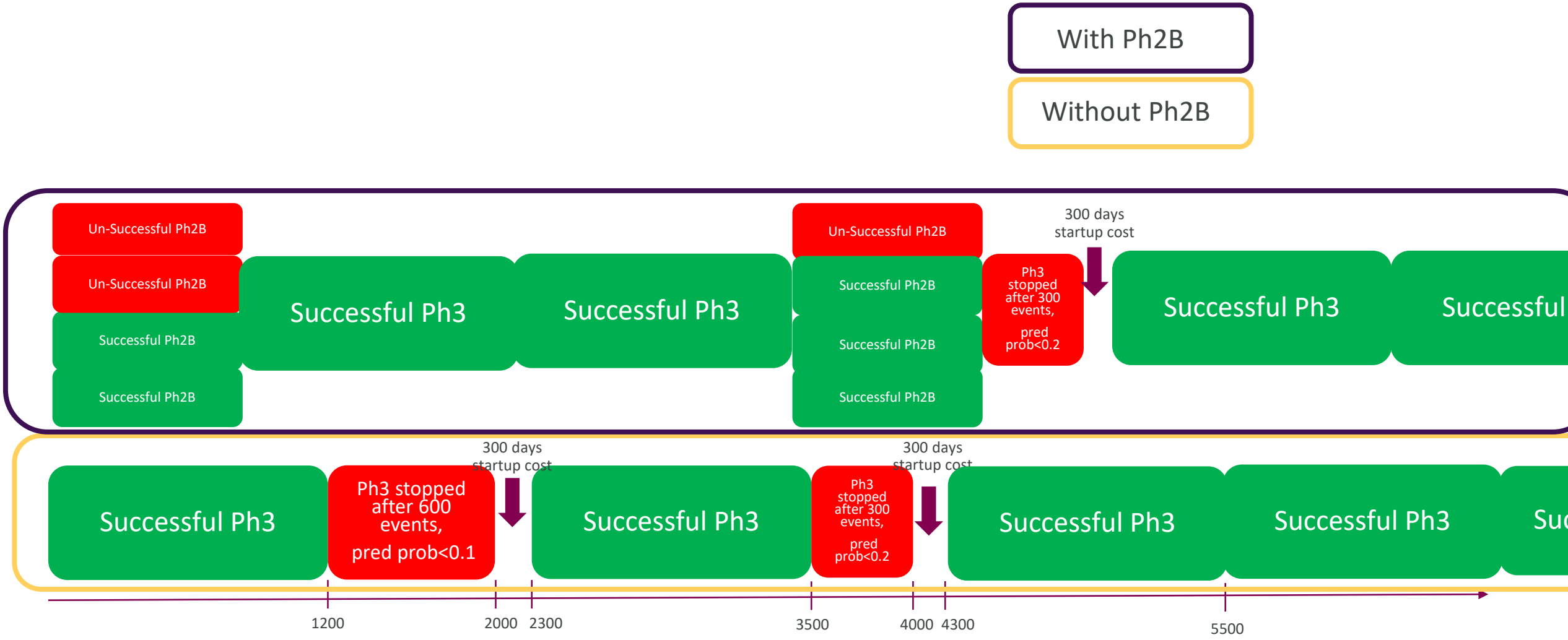
**Number of successful studies:**
3
0






Example of a run with a strategy with no Ph2B, two interims at 300 and 600 events and Gamma=(0.2,0.1)

# Illustration of the simulations with Ph2B

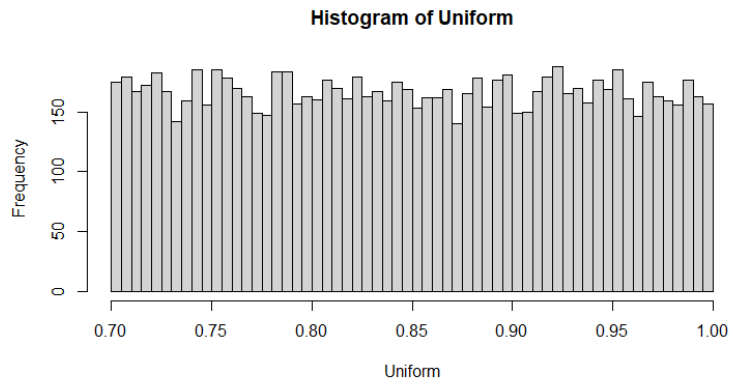


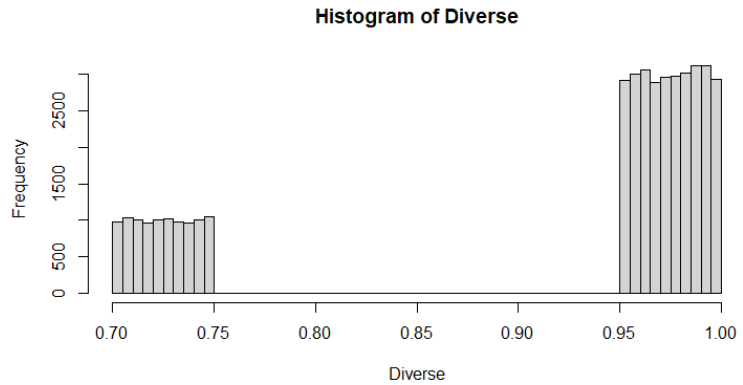

Example of a run with the same strategy as in the previous example.

# The importance of the portfolios

The distribution from which the “true” hazard ratios are simulated will greatly affect the results of the simulations. Ideally this distribution should match our portfolio of heart failure compounds. We have currently evaluated two “portfolios”:

1. Uniform between [0.7, 1]
2. Diverse, uniform between [0.7, 0.75] with p=0.25, or uniform between [0.95, 1] with p=0.75
   - The diverse portfolio was created with purpose as a discriminative portfolio, which would theoretically be an advantage for a predictive biomarker in Ph2B




The optimal strategies will differ between these two portfolios. Can we find a robust strategy that works for both portfolios? We define this as the strategy that has the largest minimal improvement across both portfolios

# Use Ph2B to predict Ph3 efficacy?

It is assumed that

- Ph2B delays the start of Ph3 by 3 years
- Daily cost of Ph2B is roughly 4.5 times smaller than Ph3 => We can do 4.5 Ph2B studies at the same time to the same daily cost

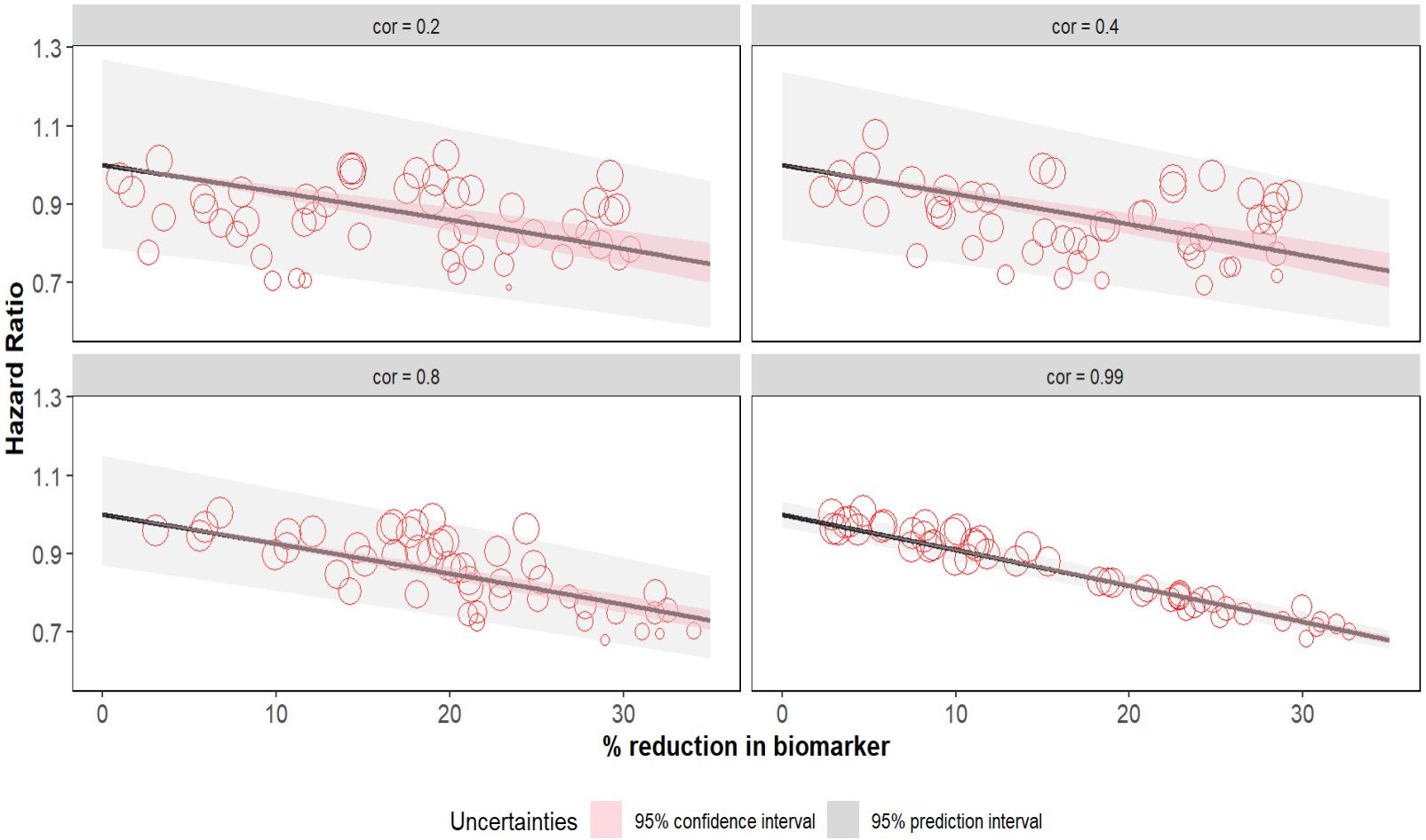
“Predictive” biomarkers in Ph2B may inform us about the expected Ph3 efficacy. Now, assume we have 170 trillion $ to spend to maximize number of successful Ph3 trials within 1 million days.

**Is it beneficial to spend some time/money on Ph2B trials to maximize # of successful studies within 1 million days, with 170 trillion $ to spend in total? If so, how predictive do Ph2B need to be?**



Detailed list of assumptions including Ph2B comparison listed in back-up

# Theoretical correlation between Ph2B and Ph3 to compute PTS in Ph3 given Ph2B




- No validated/approved Ph2 surrogate endpoint in the heart failure space
- ***Assume*** a true correlation between treatment effect on Ph2 endpoint (e.g. NT-proBNP) and Ph3 hazard ratio: 0.2; 0.4; 0.8; 0.99
- Simulate individual patient data for ~ 50 studies based on assumed correlation between endpoints (each red circle represents the summary effect in a study)
- Fit a meta-regression model to the simulated data and then use this model for prediction of HR based on treatment effects in Ph2
- Calculate the predictive probability of success (PTS) in Ph3, given Ph2 results, variability and prediction of Ph3 effect. If PTS > than a pre-specified threshold, then GO for Ph3 otherwise STOP

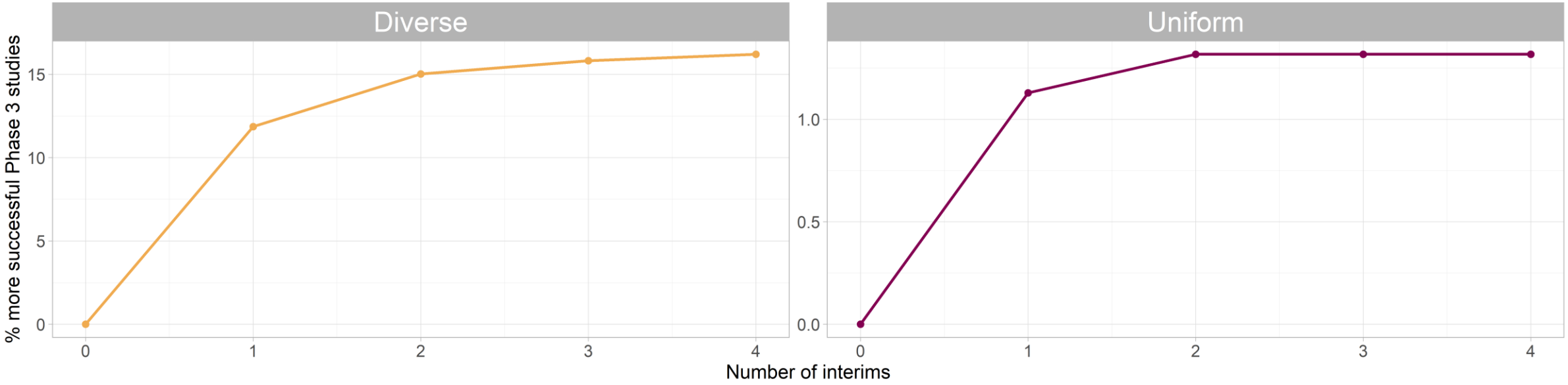
**Table. Correlations (ballpark) estimated based on the data we collected from the domain analysis⁺**

| NT-proBNP vs. HR* | LVEF vs. HR# | Domain prob(Go) vs. HR$ |
|---|---|---|
| < 0.3 | 0.2-0.4 | 0.2-0.5 |

⁺ To our knowledge, there exists no established estimate of these correlations
* Treatment effect on NT-proBNP on the log scale vs. HR on the log scale based on data from 12 compounds
# Treatment effect on LVEF on the original scale vs. HR on the log scale based on data from 10 compounds in HFrEF population
$ logit of prob(Go) estimated by domain approach vs. HR on the log scale based on data from 13 compounds

# Results

# Results for the most robust[1] strategies compared to not having any interims




| Number of interims | % more successful than no interims Uniform portfolio | % better than no interims Diverse portfolio | Robust interim positions (events) | Robust Gammas | % of studies incorrectly stopped with most robust strategy Uniform portfolio | % of studies incorrectly stopped with most robust strategy Diverse portfolio |
|---|---|---|---|---|---|---|
| 3 | 1.5% | 16% | 200,400,600 | 0.1,0.1,0.1 | ~3% | ~0.5% |

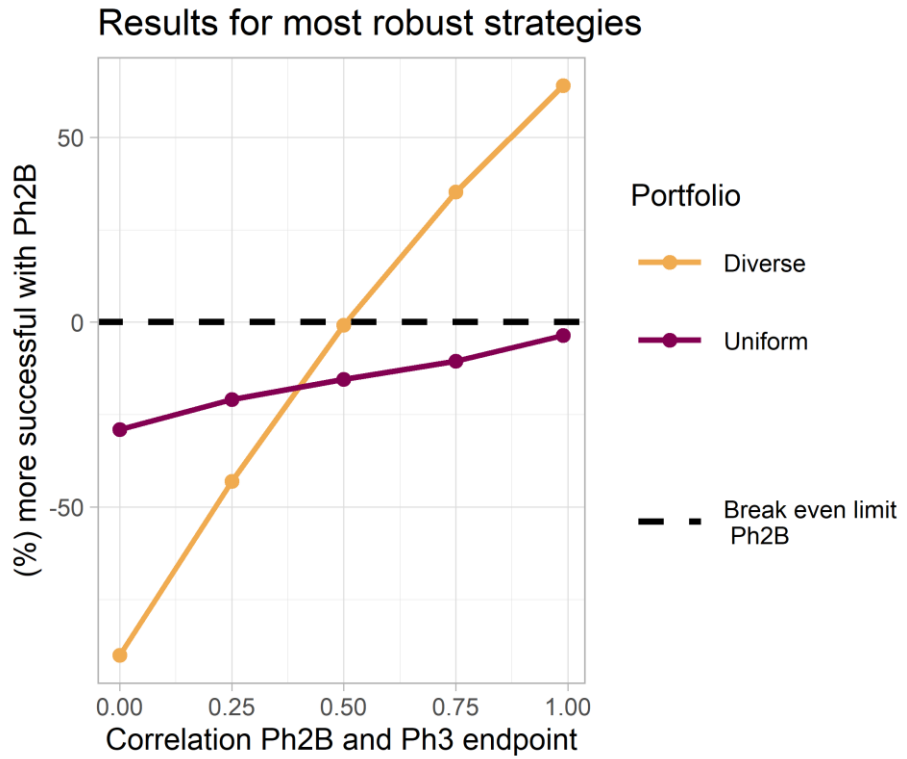
When not running a Ph2B

- Use a few (not too many) interims
- Place them quite early with low futility thresholds
- Risk of incorrect stopping using interims is 0.5-3%



[1] The strategy that has the largest minimal improvement across portfolios

# Results for the most robust[1] strategies




When comparing with and without Ph2B

- For the uniform portfolio, Ph2B is not beneficial even with a perfect biomarker
- For the diverse portfolio, correlation of 0.5 is the ”break even point” for Ph2B to be beneficial

[1] The strategy that has the largest minimal improvement across portfolios

# Conclusions

# Overall conclusions

- Currently, Ph2 endpoints in the heart failure space are unable to predict Ph3 results with high accuracy. This holds even when combining them using domains

- Number of successful Ph3 trials to the same amount of time and money could be further increased by using a futility strategy with a few (2-3) interims in the first half of the Ph3 trial, with low futility thresholds

- Given the Ph2B current cost, length and predictability of Ph2 endpoint within heart failure, number of successful Ph3 studies could be increased by running a leaner Ph2A solely focusing on dose-finding and safety followed by a Ph3

- ***"In the long run skipping Ph2B will yield at least 20% more successful Ph3 trials compared to including Ph2B, to the same cost and time"***

# Back-ups



# Back-up Domain track

# Summary of effects - methods

- Literature search strategy: systematic search in Trialtrove for selected Ph2 heart failure endpoints followed by a complementary search
- For compounds where an effect on a variable have been studied in multiple studies (e.g. 6MWD in Dapa-HFrEF population), these studies have been weighted together by the inverse of the variance (see the Cochrane handbook)
- Studies with limitations (e.g., no control group) have been manually downweighted

# Simulation set-up and assumptions for the domain-track

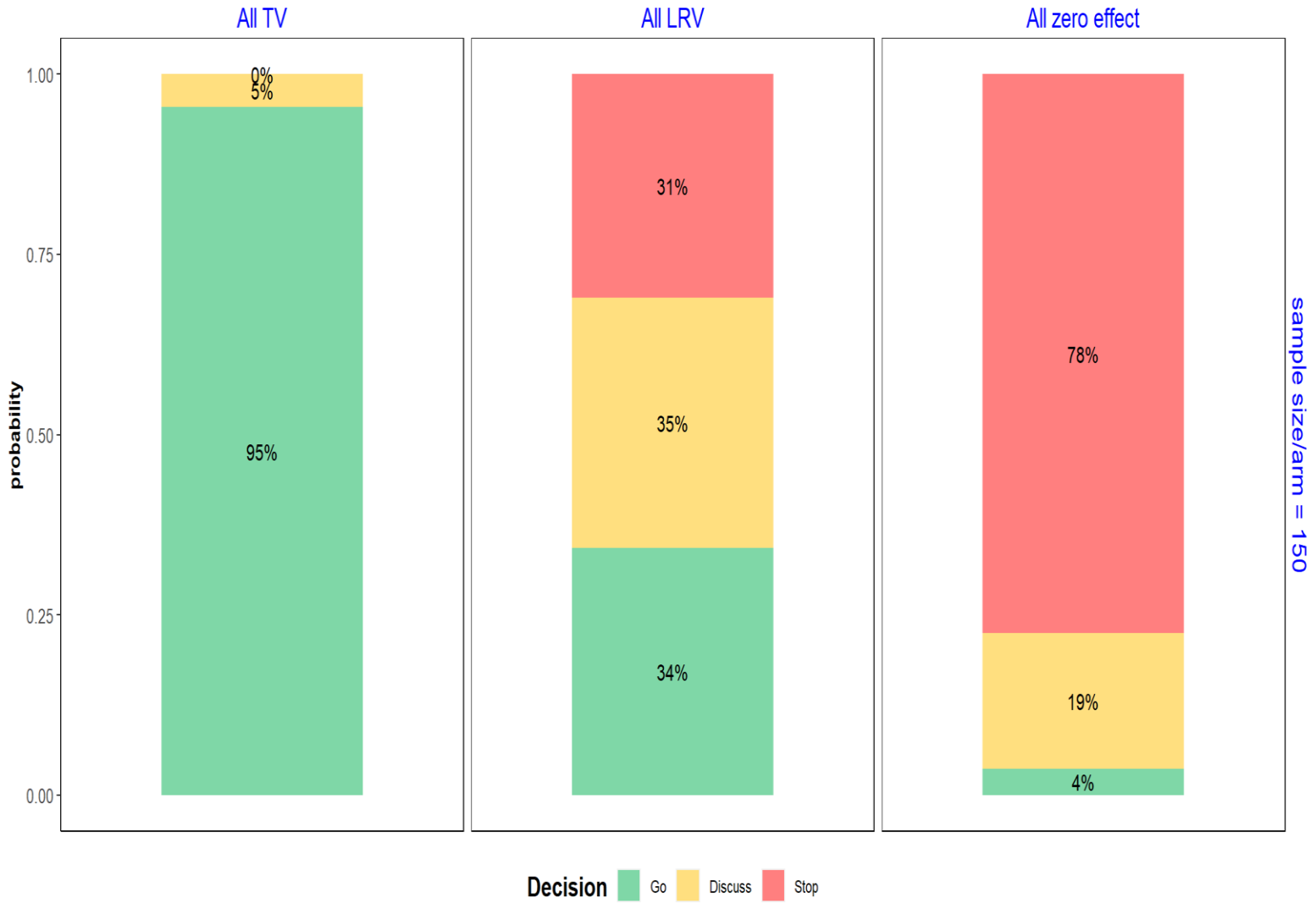
- Number of simulated trials: 1000
- Evaluable # patients/arm: 150 (also experimented with 100/arm, 200/arm, and 400/arm)
- Correlations:
  - Within domain: 0.3
  - Between domain: 0.15

  Also experimented with other assumed correlations (results are fairly robust)
- Scenarios
  - “Observed effects” scenario: applied GNG framework on all compounds data
  - “Theoretical” scenario:
    - True effects: 1) TV for all variables, 2) LRV for all variables, 3) zero treatment effect for all variables



# “Theoretical” scenarios results – interpretation (150/arm)

Under each “theoretical” scenario (all TV, all LRV, all 0 effect), we have simulated 1000 Ph2B trials, each with 150 patients/arm. We then apply the domain GNG decision framework to each simulated Ph2B trial, the probabilities of getting Go, Discuss , Stop among 1000 simulated trials are presented in the figure below:



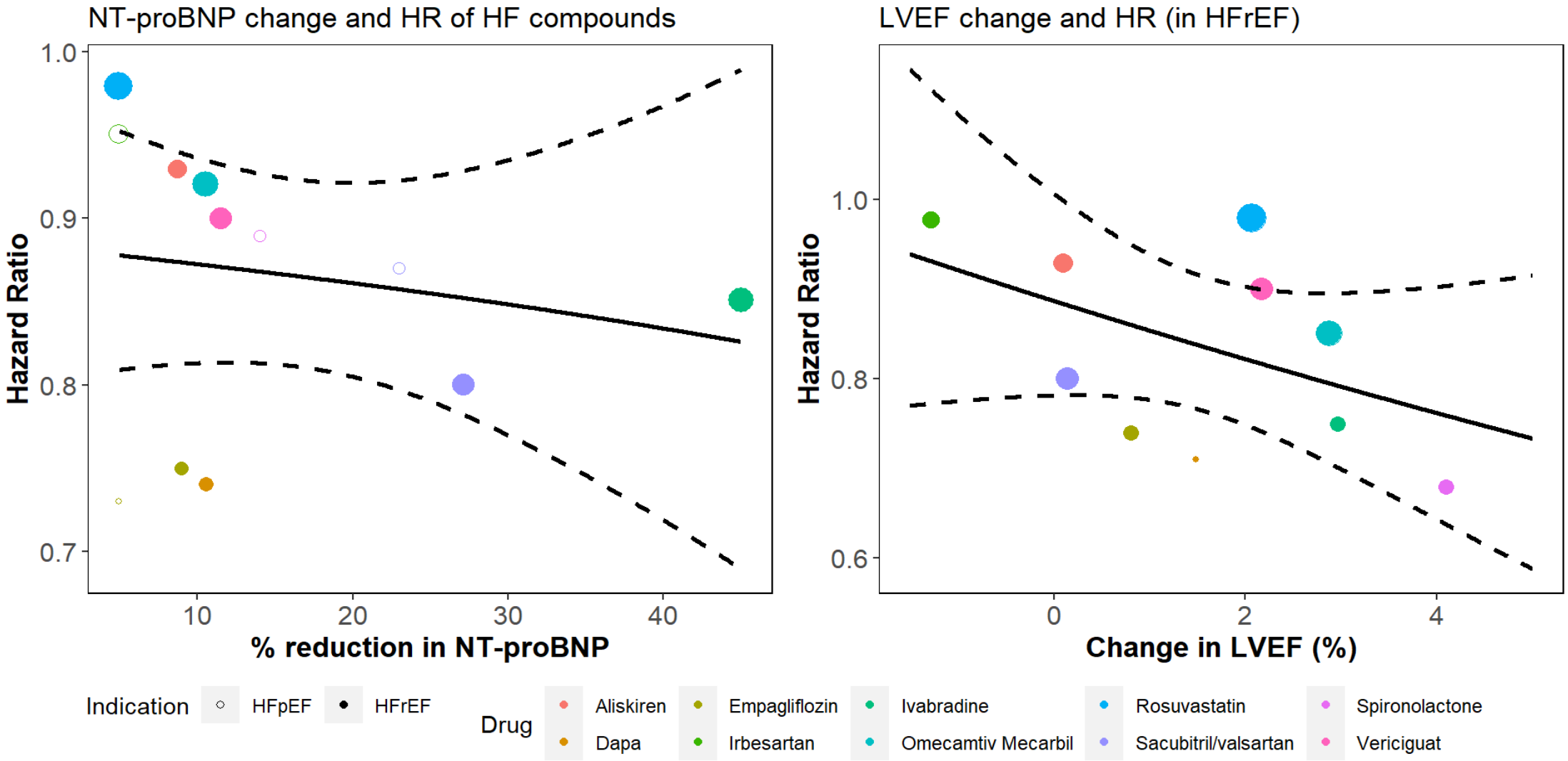

- When true treatment effect for all endpoints is their respective TV, prob(GO) is high
- When true treatment effect for all endpoints is their respective LRV, prob(GO), prob(DISCUSS), and prob(STOP) are similar, and it’s challenging to make a GNG decision based on the domain results
- When there is no treatment effect on any endpoint, prob(GO) is low

# Explorations on potential modification of the decision criteria

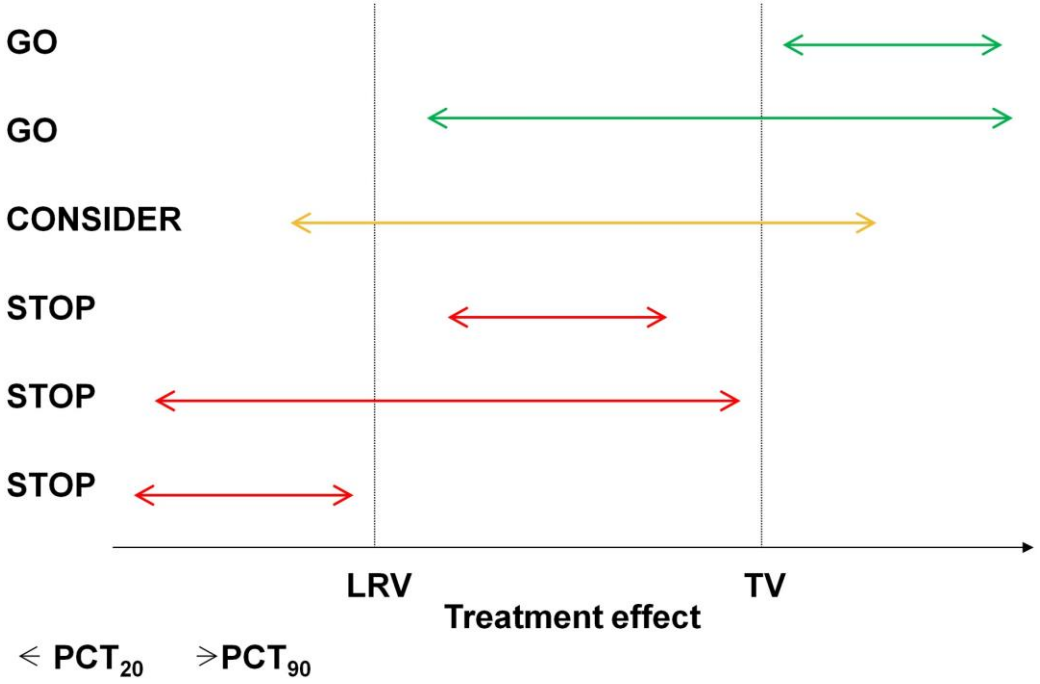
1. Use a new GO criteria for each individual endpoint to address the problem of “getting wrong decisions” under the “All LRV” scenarios
   - Improved properties under “theoretical scenarios”
   - Under “observed effects” scenarios, new criteria only makes a marginal difference to the overall results compared to original GO criteria (AZ standard)

2. Scaled score-based approach (effects/desired effect)
   - Performance similar to the domain approach (“optimised” threshold) under “theoretical scenarios”
   - Unable to test under “observed” scenarios due to high missing rate in compound data and no basis for imputation

# Meta-regression plots



# Visualization of the decision-making framework




Visualization of the decision marking framework. Frewer P., Mitchell P., Watkins C. Matcham J. *Decision-making in early clinical drug development.* (2016). The Journal of Applied Statistics in the Pharmaceutrical Industry

- The visualization shows confidence intervals in relation to the Target Value and Lower Reference Values.
- As seen in the Figure, Both the upper and lower bounds of the intervals impact the decisions.
- **GO if:** Probability of being worse than LRV<20%
- **STOP if:** Probability of being better than TV < 10% (STOP overrides GO)
- **Consider if:** Probalility of being worse than LRV>20 % and Probability of being better than TV>10%

# Back-up Bayesian futility track

# Assumptions specific for the comparison with Ph2B

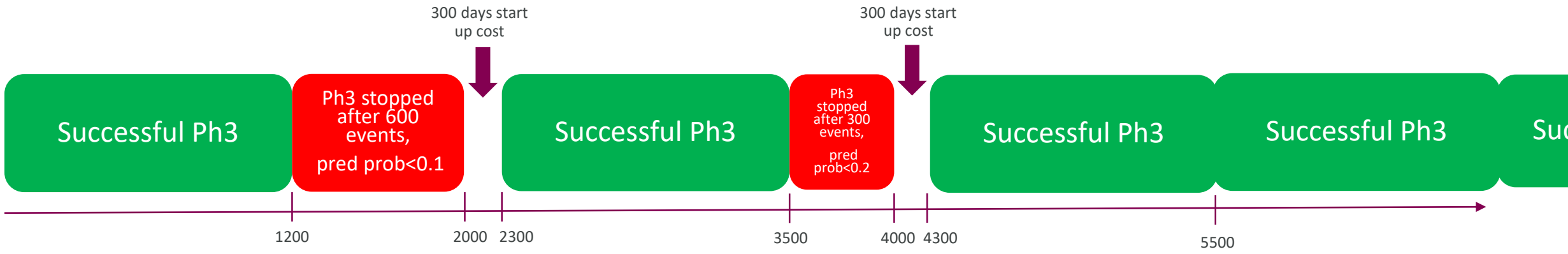
| Measure | Assumption | Explanation of assumption |
|---|---|---|
| Ph3 time | 1200 days | Ph2B and Ph3 daily cost can be summarized with that: You can run 4.5 Ph2B simultaneously to the same daily cost as a Ph3 daily cost.<br><br>Converting this to time, the simulations assumes a Ph2B takes 1095/4.5 = 243 days to run. |
| Ph3 cost | 200M USD | |
| Ph3 size | Event-driven, 1200 events, 2 arms (active vs control), 4000 patients | |
| Ph2B time | 3 years = 1095 days (delay compared to no Ph2B) | |
| Ph2B cost | 40 MUSD | |
| Ph2 size | 200 patients / arm | |
| Design features for Ph2B endpoint | SD = 0.8 (log-scale), effect size = -log(0.8) | Ph2 endpoint is assumed to be normally distributed, adjusted to achieve 80% power for current assumed Ph2B size |
| Relationship Ph2 endpoint to Ph3 endpoint | Different correlations and relationship from theoretical meta-analysis, see separate slide | |
| Evaluated predictive probability thresholds for Ph2B GNG criteria | 0.5, 0.6, 0.7 | If predictive probability of success in Ph3 given Ph2 data and the theoretical meta-analysis is larger than the pre-specified threshold, then GO for Ph3 otherwise STOP |
| Ph2B dose-finding / population-finding advantage | With 10% probability, skipping Ph2B get a punishment by adding 0.05 to the true HR. True HR can never be >1. | |
| Ph2B safety finding advantage | In 10% of the cases, drug is stopped due to safety after Ph2B.<br>If Ph2B is skipped, in 10% of the cases, drug is stopped due to safety after 300 days in Ph3. | In 10% of the cases, a compound is stopped for safety reasons after Ph2B. For Ph2B it means a punishment of the time it takes to run Ph2B (currently 243 days). Without Ph2B, this is discovered in Ph3 after 300 days, and as usual there is another 300 days before start of the next trial. So the punishment for Ph3 is 300 + 300 = 600 days. |

# Illustration of the simulations

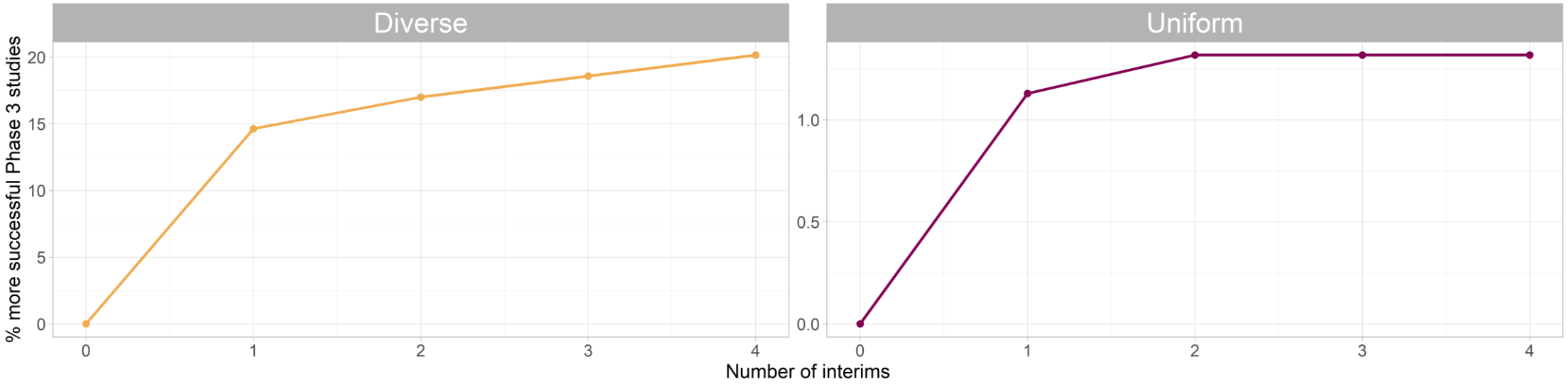
- Example of a run with a strategy with no Ph2B, two interims at 300 and 600 events and Gamma=(0.2,0.1)

Days left:0

Number of successful studies: 2



# Results for the best strategies




| Portfolio | Number of interims | % more successful than no interims | Interim positions (events) | Gammas | % of studies incorrectly stopped |
|---|---|---|---|---|---|
| Uniform | 2 | 1.5% | 400,600 | 0.1,0.1 | 2% |
| Diverse | 4 | 20% | 100,300,400,600 | 0.2,0.2,0.2,0.2 | 2% |