Strategies for improving the assessment of probability of success (PoS) in late stage drug development

Markus Lange

Joint DSBS/FMS meeting 2022, Copenhagen
22nd November, 2022

NOVARTIS | Reimagining Medicine

Acknowledgements

- **Novartis PoS team**
- **Our academic collaborators:** John Paul Gosling, Anthony O'Hagan

Hampson LV, Bornkamp B, Holzhauer B, et al. *Pharmaceutical Statistics,* 2022; 21:439

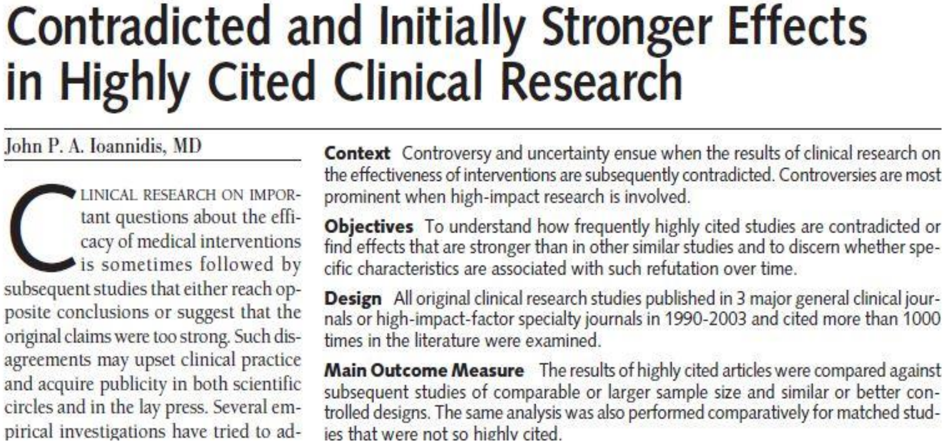
Holzhauer B, Hampson LV, Gosling JP, et al. *Pharmaceutical Statistics,* 2022. In press

Hampson LV, Holzhauer B, Bornkamp et al. *Clinical Pharmacology & Therapeutics* 2022; 111:1050

Outline

1. Motivation and background
2. Overview of PoS framework
3. Key steps in the PoS assessment
4. Eliciting expert elicitation
5. Conclusions

Promising results in smaller (early phase) trials are not always replicated by subsequent studies



Contradicted and Initially Stronger Effects in Highly Cited Clinical Research

John P. A. Ioannidis, MD

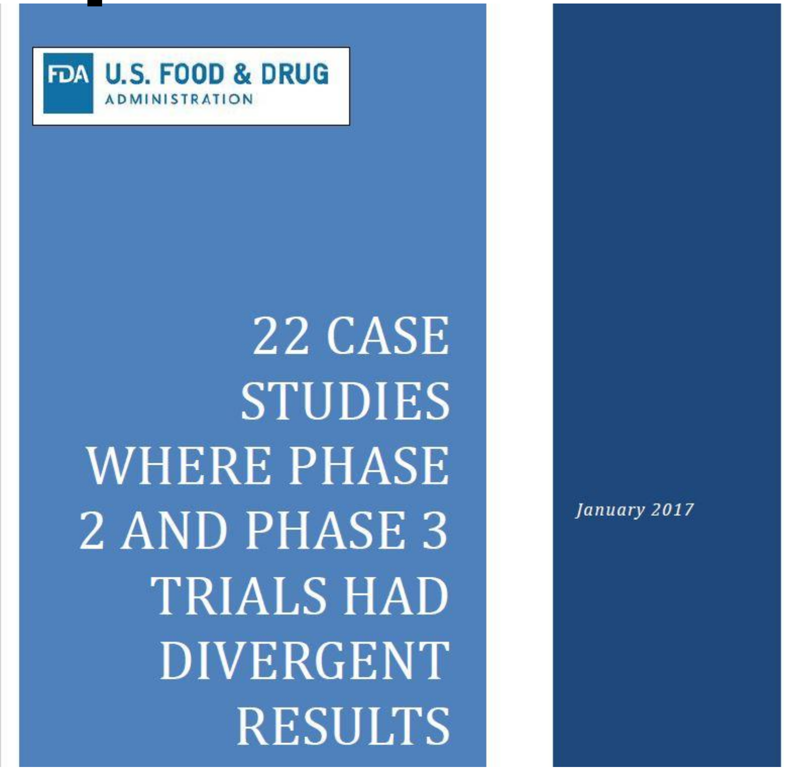
CLINICAL RESEARCH ON IMPORTANT questions about the efficacy of medical interventions is sometimes followed by subsequent studies that either reach opposite conclusions or suggest that the original claims were too strong. Such disagreements may upset clinical practice and acquire publicity in both scientific circles and in the lay press. Several empirical investigations have tried to ad-

Context Controversy and uncertainty ensue when the results of clinical research on the effectiveness of interventions are subsequently contradicted. Controversies are most prominent when high-impact research is involved.

Objectives To understand how frequently highly cited studies are contradicted or find effects that are stronger than in other similar studies and to discern whether specific characteristics are associated with such refutation over time.

Design All original clinical research studies published in 3 major general clinical journals or high-impact-factor specialty journals in 1990-2003 and cited more than 1000 times in the literature were examined.

Main Outcome Measure The results of highly cited articles were compared against subsequent studies of comparable or larger sample size and similar or better controlled designs. The same analysis was also performed comparatively for matched studies that were not so highly cited.



FDA U.S. FOOD & DRUG ADMINISTRATION

22 CASE STUDIES WHERE PHASE 2 AND PHASE 3 TRIALS HAD DIVERGENT RESULTS

January 2017

What is “success” in Probability of Success (PoS)?

- PoS is a metric quantifying the risk associated with key drug development decisions.
- PoS accounts for our uncertainty about the (unknown) effect of a drug in a Bayesian framework.
- We can calculate the PoS of a development program or an individual trial:
  - **Trial level:** Success is when a trial meets its statistical success criteria.

Probability of trial success (assurance)

Assurance is typically defined as the expected power of a trial, taking averages over a prior for the treatment effect:

$$\int \Pr(\text{Reject } H_0|\theta)\, \pi_0(\theta)\, d\theta$$

Assurance has been discussed in the following contexts:

- Choice of prior for the treatment effect: E.g. GSK base priors on elicited expert opinion.
- To inform trial design: E.g. Sample size determination; dose choice or design of a futility interim.
- To inform Ph3 go/no-go decisions
- Updating assurance after Phase 3 interim analysis

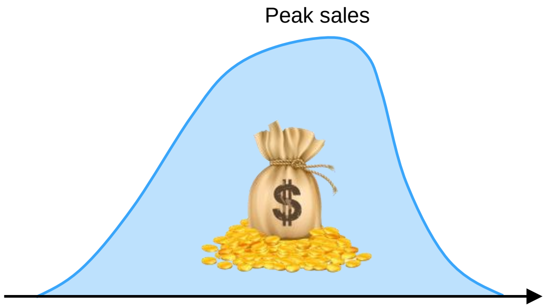
What is “success” in Probability of Success (PoS)?

- PoS is a metric quantifying the risk associated with key drug development decisions.
- PoS accounts for our uncertainty about the (unknown) effect of a drug in a Bayesian framework.
- We can calculate the PoS of a development program or an individual trial:
  - **Trial level:** Success is when a trial meets its statistical success criteria.
  - **Program level:** Success is when a program achieves regulatory approval with key endpoints needed for market access in line with their target product profile (TPP).

“Success” is more than approval: We must also meet key endpoints of the TPP required for market access

END OF PH2 — Launch — End

Expected commercial value = **Probability of “Success”** X **Potential value of “Success”**



Probability of Success within each Phase: 60% → 95% → 70%

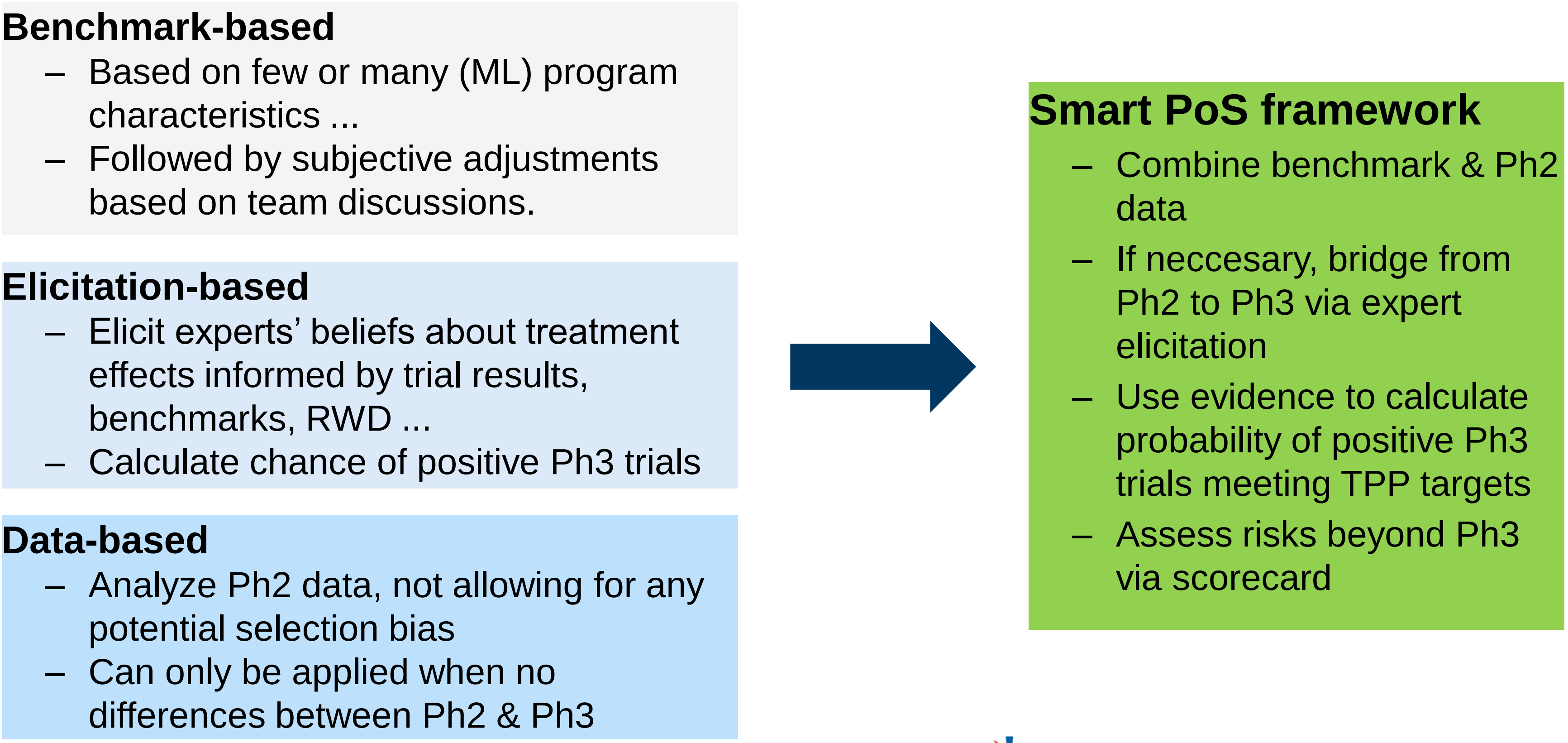
Probability of Approval only ≠ PoS: 60% * 95% = **57%** (potential gap)

Probability of Approval & TPP: 60% * 95% * 70% = **40%**

Sales forecasts assume that key endpoints of TPP are met

“Success” = Regulatory approval with key endpoints of TPP required for market access

Three of many ways to evaluate PoS



Benchmark-based

- Based on few or many (ML) program characteristics ...
- Followed by subjective adjustments based on team discussions.

Elicitation-based

- Elicit experts’ beliefs about treatment effects informed by trial results, benchmarks, RWD ...
- Calculate chance of positive Ph3 trials

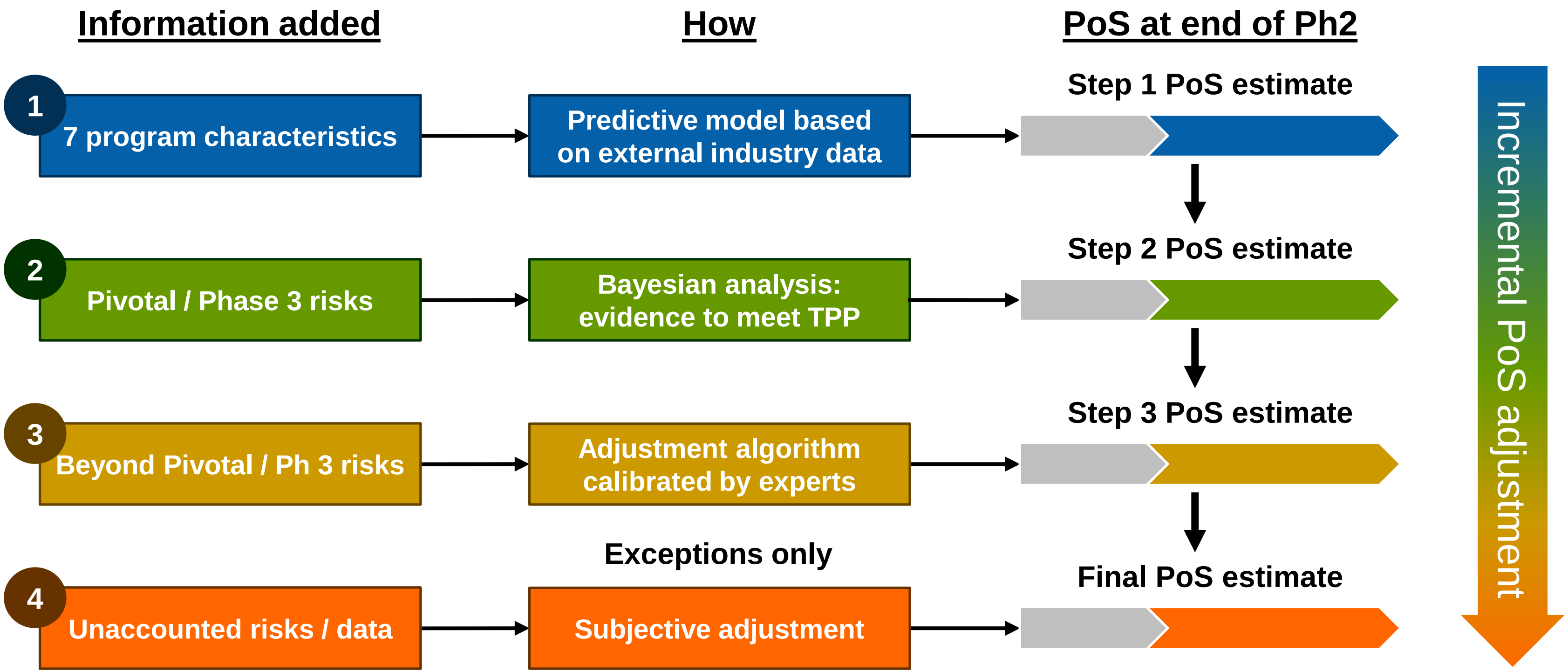
Data-based

- Analyze Ph2 data, not allowing for any potential selection bias
- Can only be applied when no differences between Ph2 & Ph3

Smart PoS framework

- Combine benchmark & Ph2 data
- If neccesary, bridge from Ph2 to Ph3 via expert elicitation
- Use evidence to calculate probability of positive Ph3 trials meeting TPP targets
- Assess risks beyond Ph3 via scorecard

How we assess PoS at the end of Phase 2 by evaluating all key evidence in 4 incremental steps



Information added | How | PoS at end of Ph2

1 7 program characteristics → Predictive model based on external industry data → Step 1 PoS estimate

2 Pivotal / Phase 3 risks → Bayesian analysis: evidence to meet TPP → Step 2 PoS estimate

3 Beyond Pivotal / Ph 3 risks → Adjustment algorithm calibrated by experts → Step 3 PoS estimate

Exceptions only

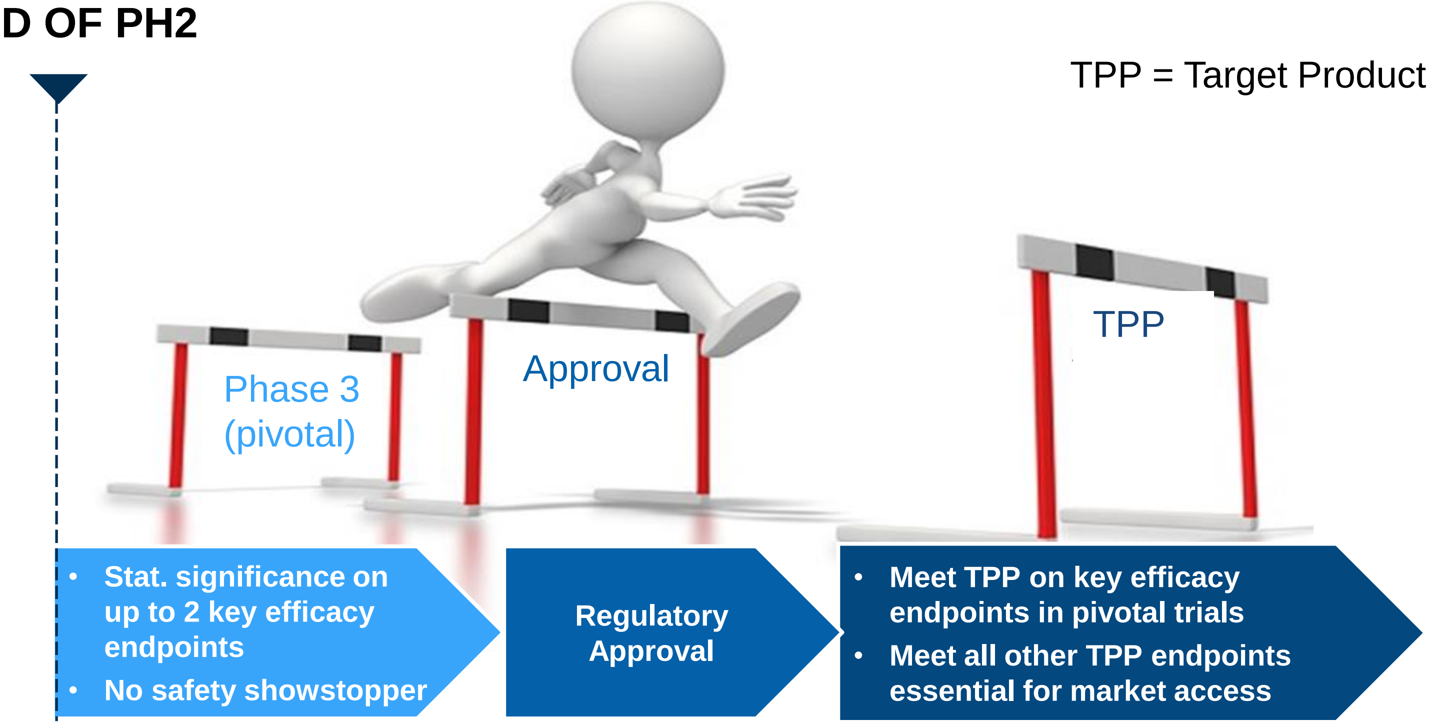
4 Unaccounted risks / data → Subjective adjustment → Final PoS estimate

Incremental PoS adjustment

Key steps in the PoS evaluation

Recap: "Success" is regulatory approval with key endpoints of TPP required for market access

TPP = Target Product Profile



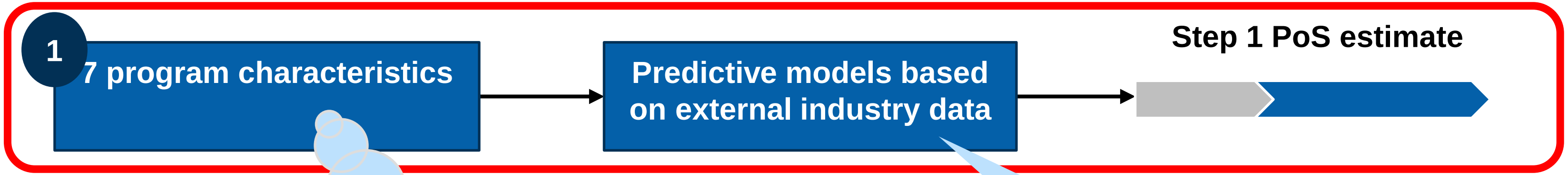
<u>What success means</u>

- Stat. significance on up to 2 key efficacy endpoints
- No safety showstopper

Regulatory Approval

- Meet TPP on key efficacy endpoints in pivotal trials
- Meet all other TPP endpoints essential for market access

Step 1: Use industry data to derive tailored benchmark for probability of approval at end of Ph2

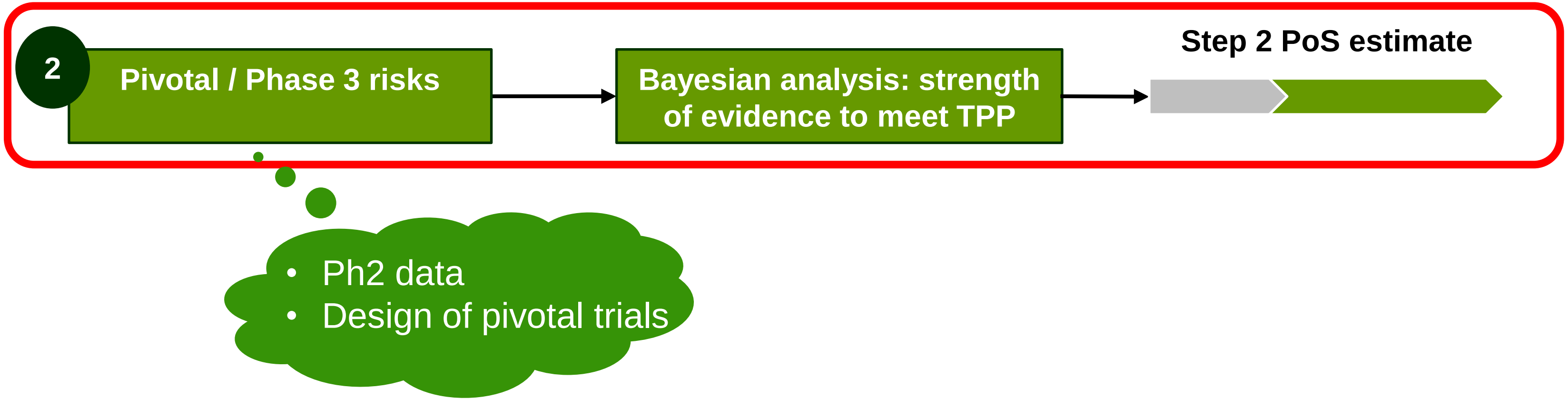

1
7 program characteristics
Predictive models based on external industry data
Step 1 PoS estimate

- **Disease Area** (11 categories)
- **Lifecycle Class** (NME / LCM / Biosimilar)
- **Molecule Class** (Protein / Small molecule / Other)
- **Drug Target** (Receptor / Enzyme / Other)
- **Route of Administration** (IV / IM / SQ / Other)
- **Size of Sponsor** (Big Pharma / Other)
- **Breakthrough Status** (Yes / No)

Considered:

- **Logistic regression**
- Lasso
- Random forest
- Neural network
- Support Vector Machine

Step 2: Leverage clinical data to assess the chance of success in pivotal studies



2

Pivotal / Phase 3 risks

Bayesian analysis: strength of evidence to meet TPP

Step 2 PoS estimate

- Ph2 data
- Design of pivotal trials

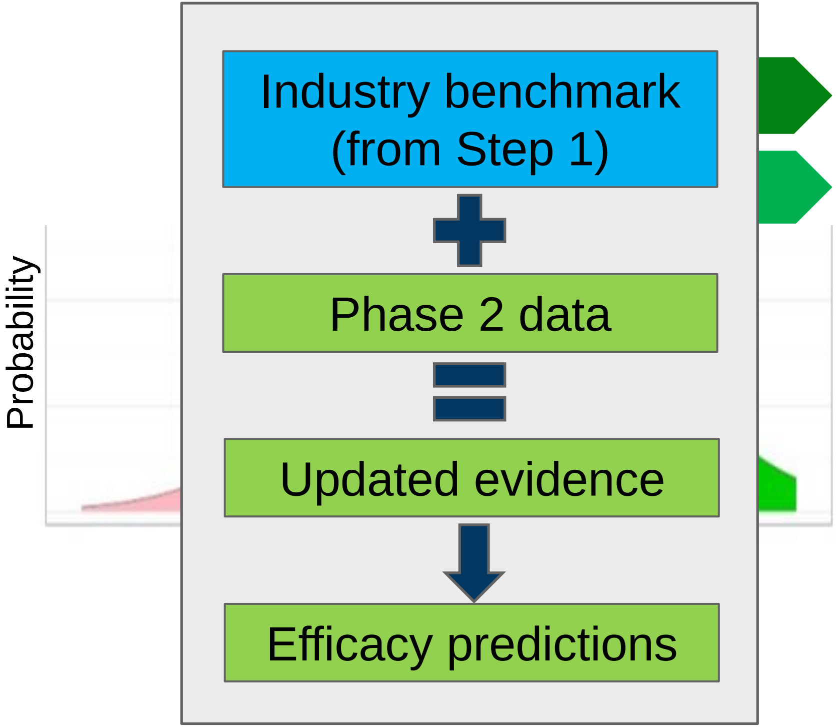
Combine external and project-specific data to assess the chance of success in pivotal trials

- Use a Bayesian approach to quantify evidence at end of Ph2 about treatment effects on 1-2 efficacy endpoints.

- Then simulate future pivotal trial(s)

- ... and assess the probability of meeting key efficacy success criteria.

- Probability of no safety showstopper is based on industry benchmark and historical reasons for failure in Ph3.



Industry benchmark (from Step 1)

+

Phase 2 data

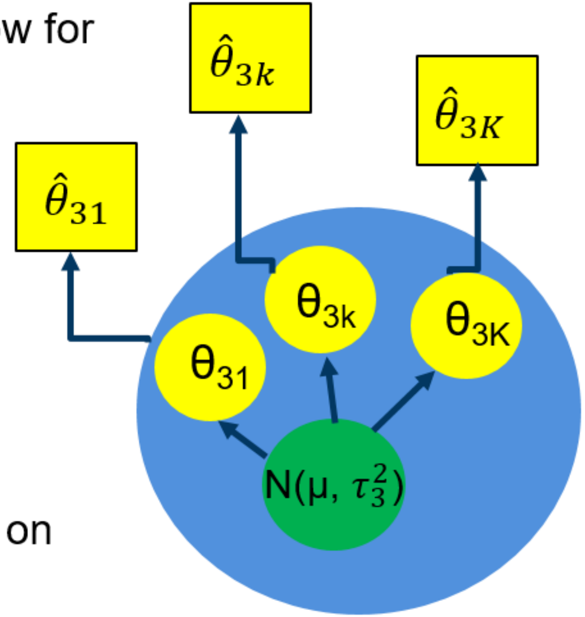
=

Updated evidence

Efficacy predictions

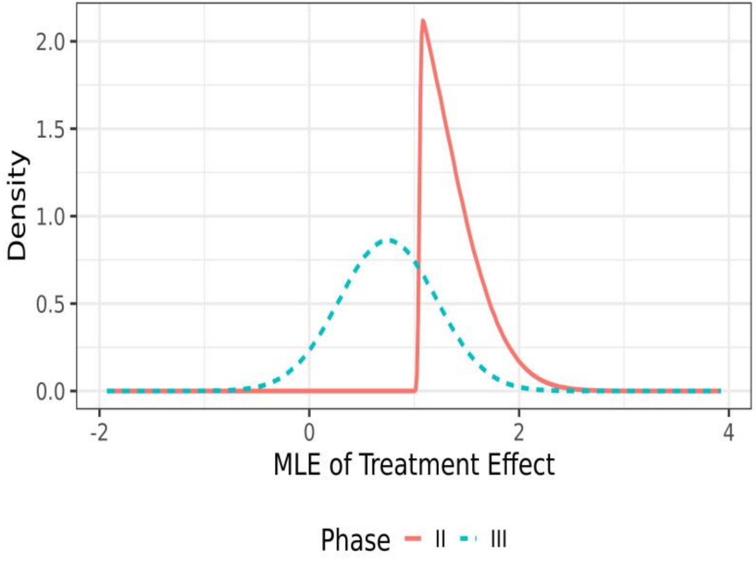
Account for between-trial heterogeneity in PoS calculation

1. Analyze: We observe effect estimates $\hat{\theta}_{2j}$ (j=1,...,J) from the Ph2 program. Fit a meta-analytic model with prior τ_2 ~ HN(z_2^2) and a prior for μ motivated by benchmark data => draw samples from posterior for μ
2. Extrapolate: Assume $\theta_{31}, \ldots, \theta_{3K} |$ μ, τ_3 ~ N(μ, τ_3^2) to allow for different between-study heterogeneity in Ph3
3. Predict (repeat *m* times):
   a) Take samples from the posterior of μ => μ* and the HN(z_3^2) prior for τ_3 => τ_3^*
   b) Take K independent samples from random-effects distribution N(μ*, τ_3^{*2}) => $\theta_{31}^*, \ldots, \theta_{3K}^*$
   c) Simulate a Phase 3 program for each sample (given the treatment effects θ_{3k}^* and Ph3 design)
4. Calulate predictive probability of efficacy success based on definition applied to each of the *m* programs



Selection bias in Phase 2 effect estimates

If we progress to a pivotal trial only if we see a promising effect in Ph2 data, we will likely see some regression towards the mean in pivotal studies.



Several possible solutions:

- Model the selection process
- Discount the Ph2 effect estimate
- Analyze Ph2 data using 'Lump and Smear' prior

Choose prior for the average treatment effect μ to ameliorate impact of potential selection bias

Problem: We want a prior for μ satisfying the following requirements:

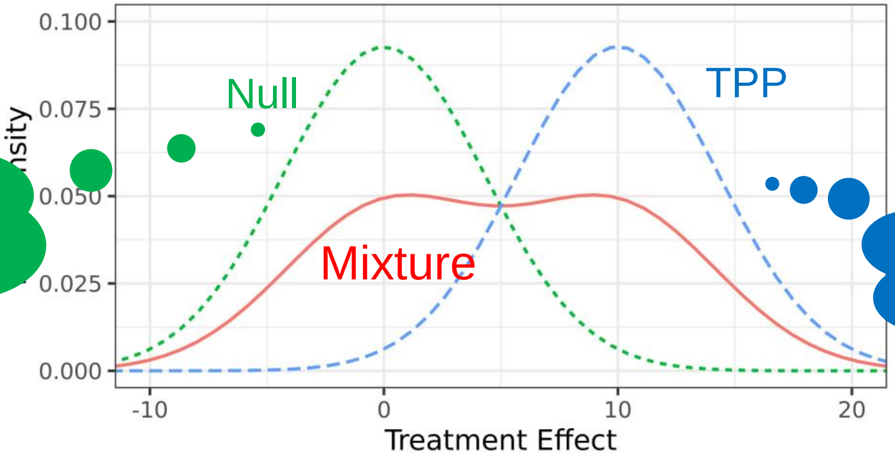
1. Prior should reflect some degree of skepticism
2. The degree of skepticism should be informed by historical success rates of similar projects at same stage of development
3. Impact of any shrinkage on the posterior should decrease as the Ph2 sample size increases and as the efficacy signal increases.

Solution: We use a mixture prior for μ with weights calibrated to industry benchmark chance of efficacy success in Ph2 and pivotal trials.

Specify prior for average effect μ which is mixture of two normal distributions

Null component

- Mean = null
- P(μ > TPP) = 0.01



TPP component

- Mean = TPP
- P(μ < 0) = 0.01

Mixture Prior: $w_N*N(0, \sigma_N^2) + (1 - w_N)*N(TPP, \sigma_T^2)$

- Calibrate w_N to ensure the marginal probability of a ‘standard Ph2 & Ph3 program’ succeeding equals the industry benchmark chance of efficacy success in Ph2 & Ph3.

Simulate future pivotal studies to calculate the predictive probability of efficacy success

- We do not simulate individual patient data. Rather simulate standardized test statistics assuming that :

$$\begin{pmatrix} Z_{1i} \\ Z_{2i} \end{pmatrix} |(\theta_{1i}^*, \theta_{2i}^*) \sim N\left(\begin{bmatrix} \theta_{1i}^*\sqrt{\mathcal{J}_{1i}} \\ \theta_{2i}^*\sqrt{\mathcal{J}_{2i}} \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right)$$

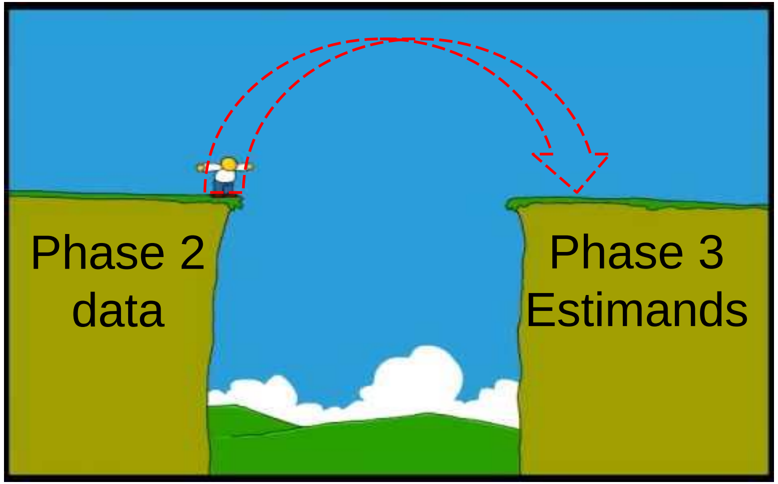
where ρ is the within-patient correlation of outcomes on the efficacy endpoints, and $\mathcal{J}_{ji}$ is Fisher's information for θ_{ji}^*.

- Estimate Pr(succeed on pivotal efficacy endpoints) by

(# simulated pivotal programs meeting success criteria)/N

Assessment of PoS is more complex when there are differences between Ph2 and Ph3

- Different phases can use different:
  - Endpoints
  - Patient populations
  - Comparator arms
  - Dose regimens
- Relate Ph2 data to pivotal quantities of interest by **eliciting expert opinion.**

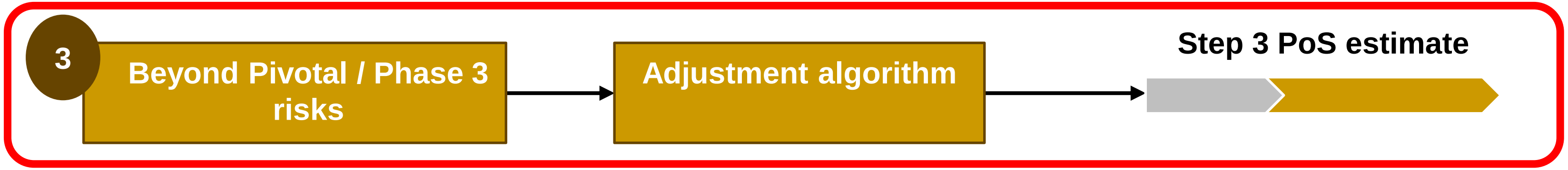


Source: Joe Cartoon

What is elicitation?

- The process of
  - representing the knowledge
  - of one or more persons (experts)
  - concerning an uncertain quantity
  - as a **probability distribution** for that quantity.
- Typically conducted as a dialogue between
  - the experts – who have substantive knowledge about the quantity of interest – and
  - a facilitator – who has expertise in the process of elicitation
  - Ideally face to face
    - but may also be done by video-conference

Step 3: Accounting for risks beyond pivotal studies



Program team fills in scorecard rating their project on 5 risks

Rate project low / medium / high risk on:

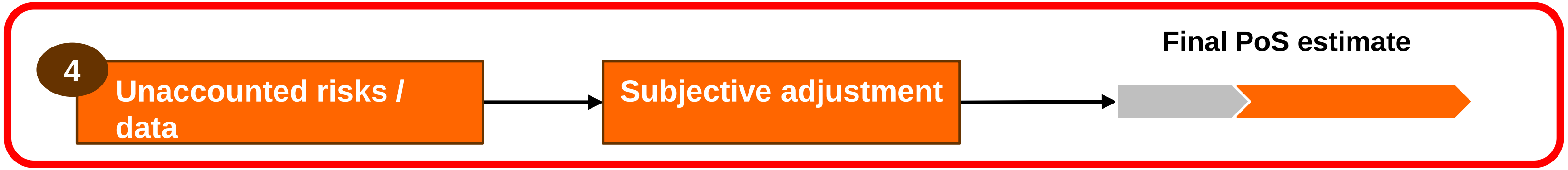
1. Alignment with key regulator
2. Unaccounted safety risks
3. Quality & compliance risks
4. Technical development risks
5. Unaccounted target product profile (TPP) risks

Examples:

1. Non-endorsed primary endpoint
2. Safety risk found in pre-clinical study
3. Inexperienced sites to be used in Ph3
4. Different inhalers used in Ph2 & Ph3
5. Additional QoL endpoint required for access unlikely to meet TPP

Benchmark chance of success in submission (from Step 1) is adjusted according to risk profile. Adjustment is based on an elicitation survey involving 30 internal experts.

Step 4: In exceptional cases, apply an adjustment in case of risks / data unaccounted for in Steps 1-3



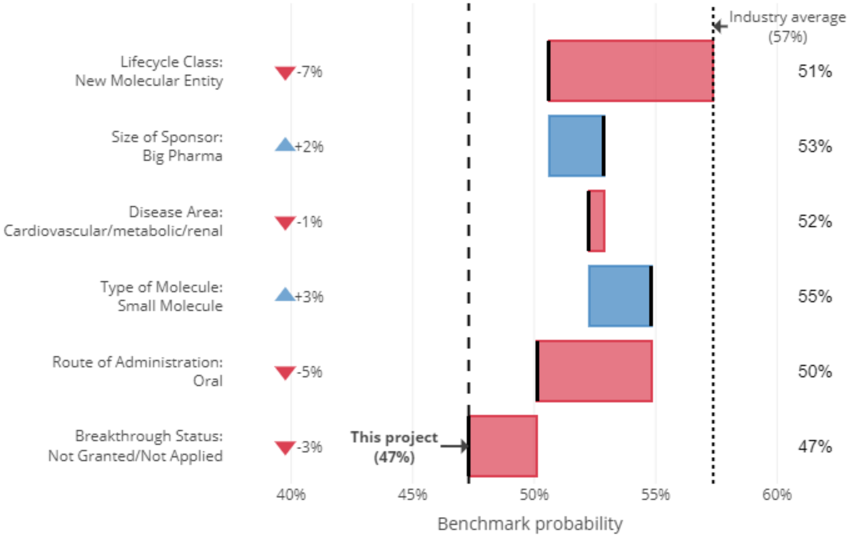
Illustrative Example

Hypothetical example

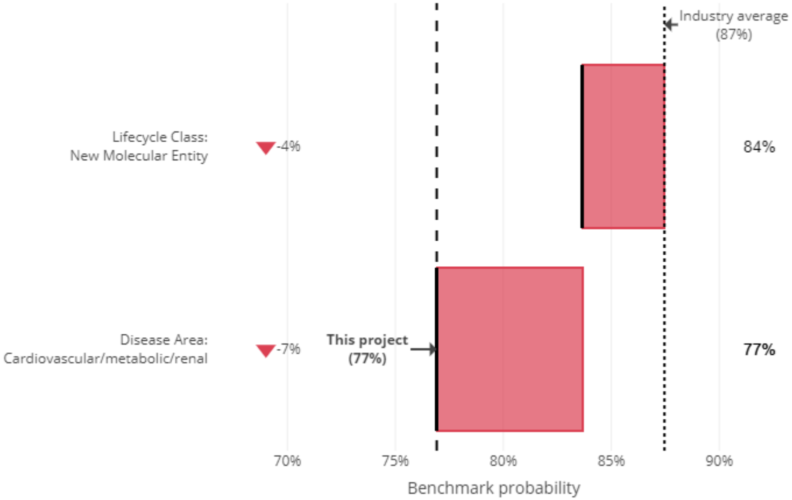
- Weight-loss drug called ThinFast
  - Small molecule, orally administered new molecular entity targeting an enzyme
  - Part of the metabolic therapeutic area
  - Health Authority has mild concerns regarding the plan to have a single Phase 3 study
- Primary Endpoint is “Weight Loss after 1 year (in kg)”
  - Used in both Phase 2b and Phase 3
  - Continuous endpoint: measured as difference in average change (vs placebo)
  - Null treatment effect: 0kg; TPP base case: 10kg
  - Standard deviation is known: 10kg
- Promising Phase 2b result: 12kg, 95%-CI: (0kg,24kg)
- One Phase 3 trial is planned
  - Sample size: 100 patients per arm
  - Testing at one-sided significance level of 0.025

Example: Step 1

Benchmark prob. of successful Ph3 = 47%



Benchmark prob. of approval after submission = 77%

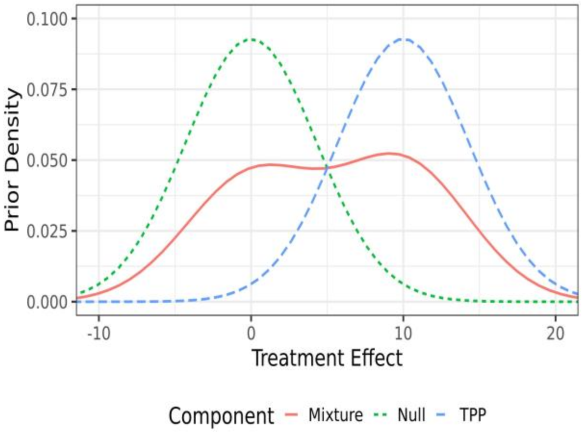


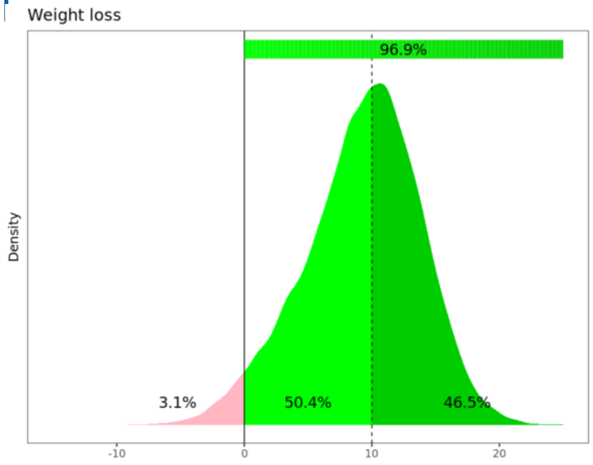
Example: Step 2

Set-up prior – Mixture prior calibrated to 32% benchmark probability of efficacy success in Ph2b & Ph3

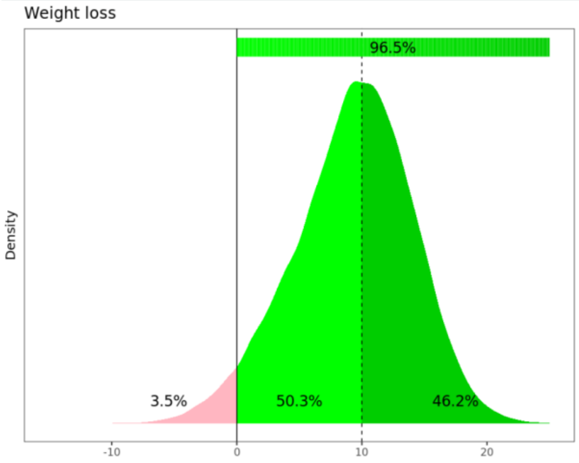
Update with Ph2b data – Derive MAP prior for treatment effect in Ph3 given Ph2b result: estimate = 12kg, 95% CI (0kg, 24kg)

Predict Ph3 – Predictive distribution for the treatment effect estimate that will be observed at the end of Ph3. Benchmark prob. of no safety showstopper is 92%





Difference in means



Effect estimate in Ph3

Example: Step 2

- Of the simulated Ph3 trials:
  - 91% achieved stat. significance on the primary endpoint
  - 84% achieved stat. significance **and** saw no safety showstopper
  - 43% achieved stat. significance **and** met the TPP **and** saw no safety showstopper

Example: Step 3

- Project was assigned the following risk ratings by the team:
  - Alignment with Key regulator: Medium
  - Unaccounted safety risks: Medium
  - Quality & compliance risks: Low
  - Technical development risks: Low
  - Unaccounted TPP risks: Low

- Given this info, Pr(Approval & remaining TPP | Pivotal Efficacy, Safety) is **61%**
- If all 5 risks had been scored as “low”, this probability would have been **84%**

Final PoS estimate

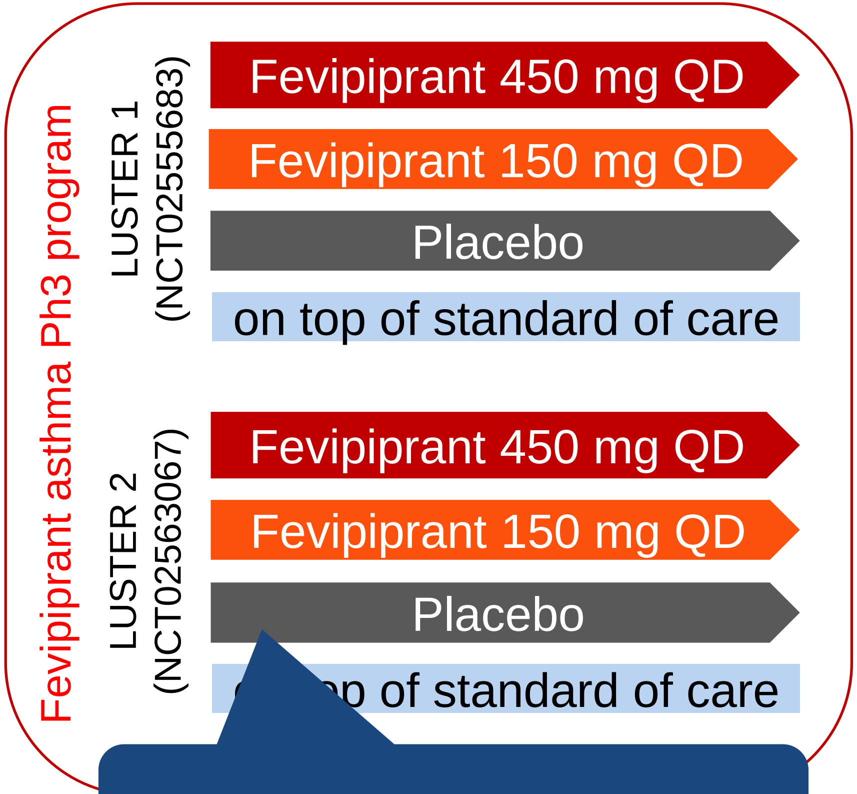
- There were no exceptional circumstances warranting a Step 4 adjustment.
- Final PoS estimate is therefore:

Probability of Success in each Phase	84%	61%	51%
Probability of Approval only ≠ PoS	84% * 61% = 51%		
Probability of Approval & TPP	84% * 61% * 51% = 26%		

Eliciting expert opinion

Example of an asthma development program

- Fevipiprant is a treatment for asthma.
- Pilot for PoS framework at Novartis
- We calculated the probability of success while the Ph3 program was underway but before DBL.
- Differences between Ph2 vs Ph3:
  - Primary endpoint: Annual rate of asthma exacerbations in Ph3
  - One Ph2 study had measured the surrogate of reduction in sputum eosinophil counts.



Fevipiprant asthma Ph3 program

LUSTER 1 (NCT02555683)

Fevipiprant 450 mg QD

Fevipiprant 150 mg QD

Placebo

on top of standard of care

LUSTER 2 (NCT02563067)

Fevipiprant 450 mg QD

Fevipiprant 150 mg QD

Placebo

on top of standard of care

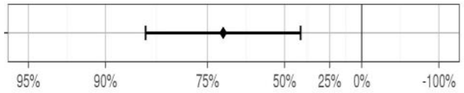
Two identical pivotal Ph3 RCTs

Using elicitation to map Ph2 data on sputum eosinophils to treatment effect on Ph3 endpoint

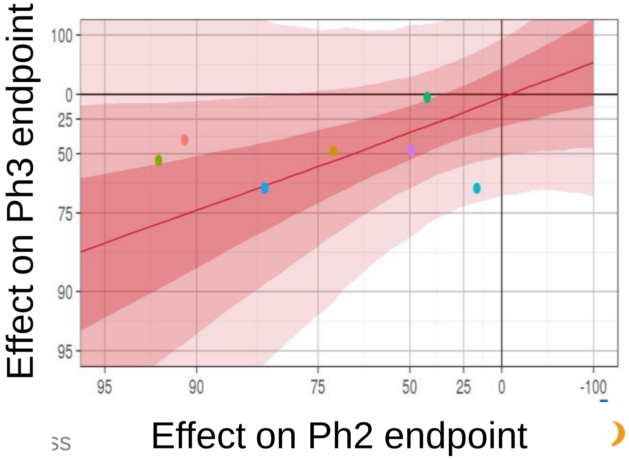
Analyze – Use Ph2 data to create a meta-analytic-predictive (MAP) prior for the treatment effect on the Ph2 endpoint in new study

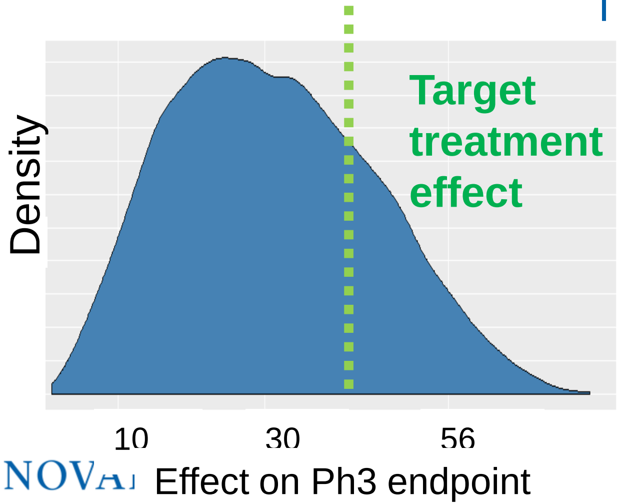
Elicit – Elicit conditional expert opinion on size of treatment effect on Ph3 endpoint under different scenarios for the size of the true effect on Ph2 endpoint

Synthesize – Use expert judgements to translate Ph2 evidence & derive marginal prior for the treatment effect on Ph3 endpoint in Ph3



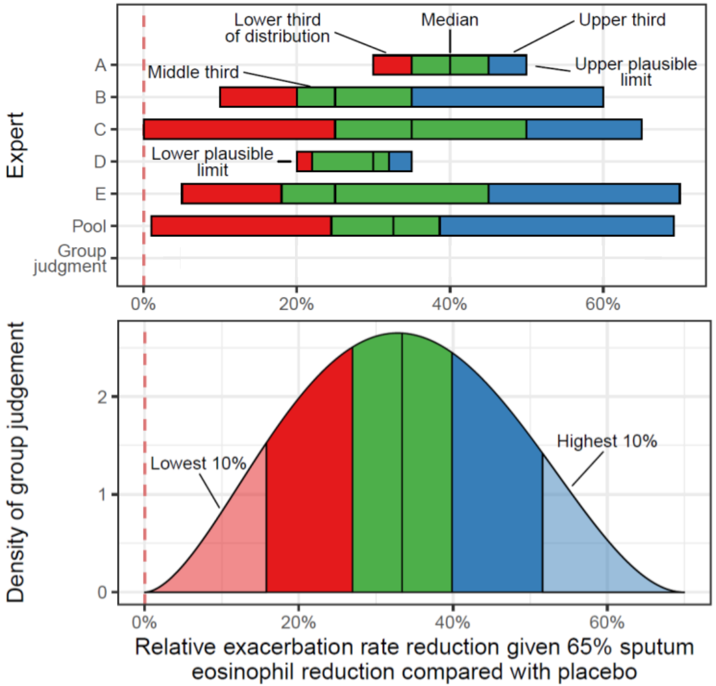
Treatment effect on Ph2 endpoint





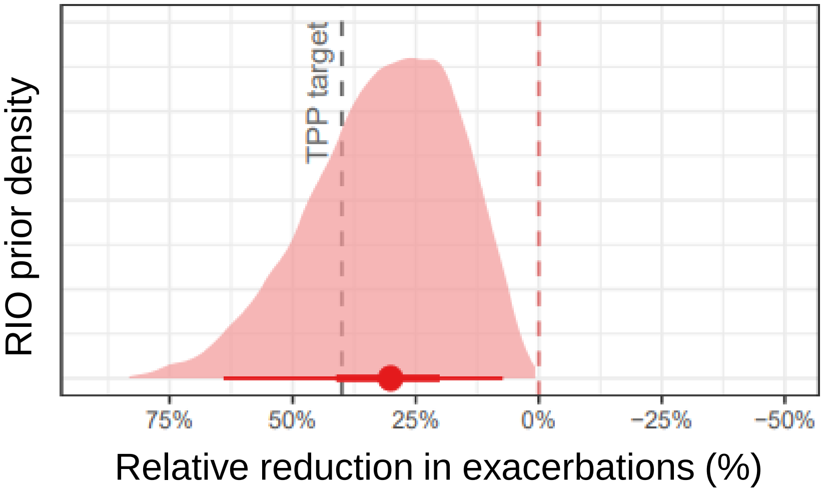
Elicitation for exacerbation rate reduction given median effect on surrogate

- Start with individual judgments
- Tertile method: in order of plausible limits, median, and then lower/upper tertile
- Each expert writes down independently
- “Challenge your judgment”
- Individual judgments revealed to group
- Group discussion
- What would RIO (a **R**ational **I**mpartial **O**bserver) think? (probability method)



Relative exacerbation rate reduction given 65% sputum eosinophil reduction compared with placebo

Comparison of RIO prior with Ph3 results

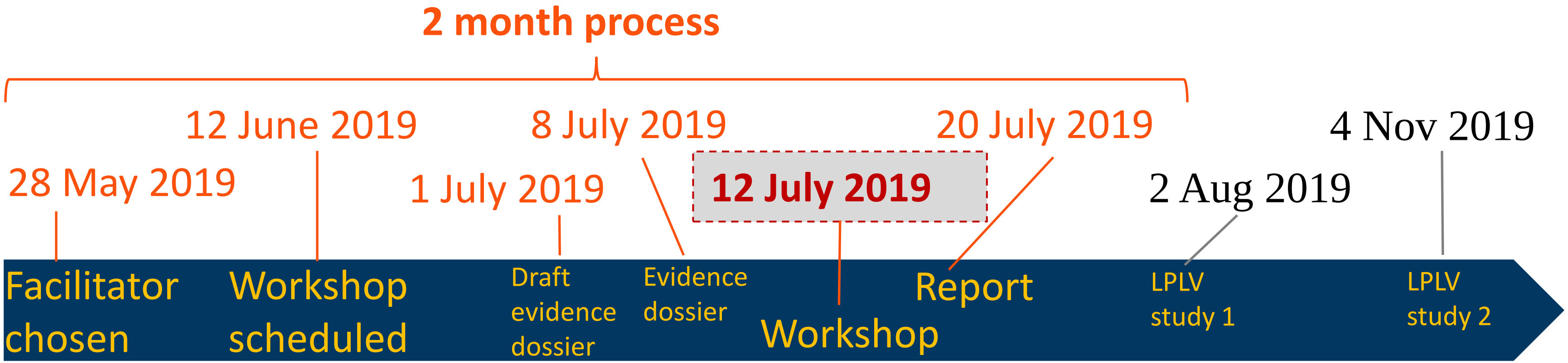


Prior median: 30.2%
95% Credible Interval: 7.0% to 60.2%

- RIO prior was consistent with the outcome of the LUSTER 1 & 2 Ph3 trials
- Observed reduction in the exacerbation rate was 23% (95% CI: 3 – 39%) based on a pooled analysis of LUSTER 1 & 2 for fevipiprant dose 450mg

A successful elicitation meeting requires careful preparation

- Defining the questions
- Identifying the relevant evidence / assembling evidence dossier
- Selection of experts



2 month process

28 May 2019 Facilitator chosen

12 June 2019 Workshop scheduled

1 July 2019 Draft evidence dossier

8 July 2019 Evidence dossier

12 July 2019 Workshop

20 July 2019 Report

2 Aug 2019 LPLV study 1

4 Nov 2019 LPLV study 2

Conclusions

Conclusions (1)

- Proposed methodology
  - ✓ Produces more reliable PoS estimates which enable better decisions
  - ✓ Increases transparency
  - ✓ Uses all available information from several sources
  - ✓ Provides insights on the impact of risk factors

- If direct data are unavailable for a QoI, expert elicitation is an attractive solution, but requires a structured process and thorough preparation

- Feedback from the experts: they found the evidence dossier a helpful resource in itself and appreciated the rigorous process and quality of the discussions

Conclusions (2)

- PoS framework is currently being implemented within Novartis
- We implemented a 2-stage roll-out
  - Worked closely with 5 early adopter teams to assess PoS at their FDP
  - After each early adopter, collected feedback to optimize process
  - Presented final process to senior management
  - After endorsement, process became mandatory as a part of wider roll-out
- Ongoing change management
  - Continue to offer trainings
  - Facilitate experience sharing
  - Ongoing refinements of methodology and processes where necessary

Bibliography

- Spiegelhalter DJ, Freedman LS. A predictive approach to selecting the size of a clinical trial based on subjective clinical opinion. Statistics in Medicine 1986; 5:1-13
- Rufibach K, Burger HU, Abt M. Bayesian predictive power: choice of prior and some recommendations for its use as probability of success in drug development. Pharmaceutical Statistics 2016; 15:438-46.
- Crisp A, Miller S, Thompson D, Best N. Practical experiences of adopting assurance as a quantitative framework to support decision making in drug development. Pharmaceutical Statistics 2018; 17;317-28
- O'Hagan A, Stevens JW, Campbell MJ. Assurance in clinical trial design. Pharmaceutical Statistics 2005; 187-201
- Dallow N, Best N, Montague TH. Better decision making in drug development through adoption of formal prior elicitation. Pharmaceutical Statistics 2018; 17:301-16
- Rufibach K, Jordan P, Abt M. Sequentially updating the likelihood of success of a Phase 3 pivotal time-to-event trial based on interim analyses or external information. Journal of Biopharmaceutical Statistics 2016; 191-201

Bibliography

- Bayarri MJ, Berger J. Robust Bayesian analysis of selection models. Annals of Statistics 1998, 26:645-59
- Kirby S, Burke J, Chuang-Stein C, Sin C. Discounting phase 2 results when planning phase 3 clinical trials. Pharmaceutical Statistics 2012; 11; 373-85
- Brightling, Christopher E., et al. "Effectiveness of fevipiprant in reducing exacerbations in patients with severe asthma (LUSTER-1 and LUSTER-2): Two phase 3 randomised controlled trials." *The Lancet Respiratory Medicine* 9.1 (2021): 43-56.
- Jeremy E. Oakley and Anthony O'Hagan. SHELF: the Sheffield Elicitation Framework (version 4). School of Mathematics and Statistics, University of Sheffield, UK, 2019. Available at http://tonyohagan.co.uk/shelf

Thank you

NOVARTIS | Reimagining Medicine