# Generalized Pairwise Comparisons: a statistical method for patient-centric medicine

*Marc Buyse, ScD*
*IDDI and U Hasselt, Belgium*

*QuanTIM Webinar*

*17 February 2023*

# Agenda

- Theory
  - Generalized Pairwise Comparisons
  - Net Treatment Benefit
- Applications
  - Augmenting power *and* clinical relevance
  - Benefit / risk analyses
  - Multiple testing procedures
- Conclusions

# Theory

# Wilcoxon rank-sum test

| TREATMENT GROUP (T) | CONTROL GROUP (C) |
| --- | --- |
| $X$  |  $Y$ |
| $X_1$ | $Y_1$ |
| $X_2$ | $Y_2$ |
| ... | ... |
| $X_n$ | $Y_m$ |

*Wilcoxon. Biometrics 1945;1:80*

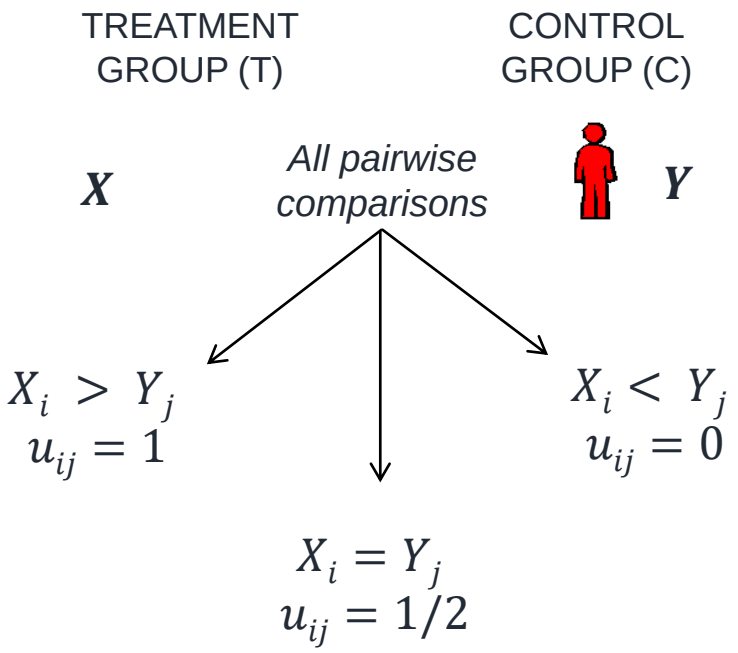
# Wilcoxon rank-sum test

| TREATMENT GROUP (T) | CONTROL GROUP (C) |
|---|---|
| $\boldsymbol{X}$  |  $\boldsymbol{Y}$ |
| $X_1$ | $Y_1$ |
| $X_2$ | $Y_2$ |
| … | … |
| $X_n$ | $Y_m$ |

1. Order the $(n+m)$ elements of $\boldsymbol{X} \cup \boldsymbol{Y}$
2. Let $R_i$ be the rank order of the $i^{\text{th}}$ element
3. For groups of tied values, assign a rank equal to the midpoint of the unadjusted ranks
4. Calculate $U = \sum_{i=1}^{n} R_i$, the sum of ranks of the elements of $\boldsymbol{X}$
5. The statistic $U$ has a known distribution under $H_0$

*Wilcoxon. Biometrics 1945;1:80*

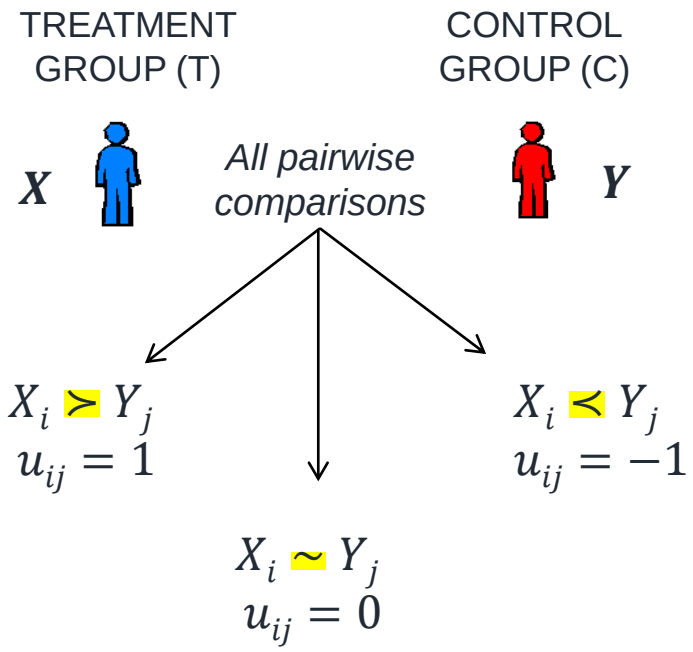
# Mann-Whitney test




1. Perform pairwise comparisons between all elements of $\boldsymbol{X}$ and $\boldsymbol{Y}$
2. Calculate $u_{ij} = \begin{cases} 1 \; if \; X_i > Y_j \\ 0 \; if \; X_i < Y_j \\ 1/2 \; \; if X_i = Y_j \end{cases}$
3. The statistic
   $U = \frac{1}{m \cdot n} \sum_{i=1}^{m} \sum_{j=1}^{n} u_{ij}$
   has a known distribution under $H_0$

*Mann & Whitney. Ann Math Stat 1947;18:50*

# Generalized Pairwise Comparisons (GPC)

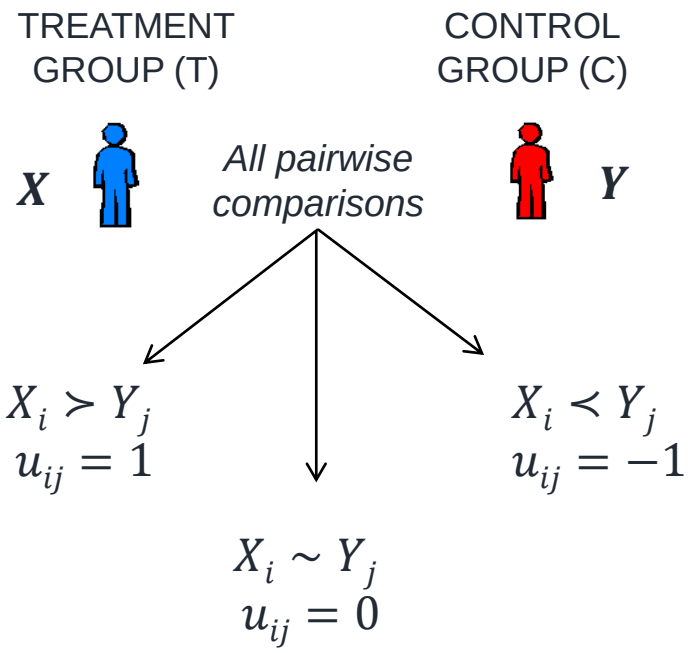



1. Perform pairwise comparisons between all elements of $\boldsymbol{X}$ and $\boldsymbol{Y}$
2. Calculate $u_{ij} = \begin{cases} +1 \; if \; X_i \succ Y_j \\ -1 \; if \; X_i \prec Y_j \\ 0 \; if \; X_i \sim Y_j \end{cases}$
3. The statistic
   $U = \frac{1}{m \cdot n} \sum_{i=1}^{m} \sum_{j=1}^{n} u_{ij}$
   has a known distribution under $H_0$

where $\succ$ stands for "better" (win)
$\prec$ stands for "worse" (loss)
$\sim$ stands for "similar" (tie) or "unclassified" (?)

*Buyse. Stat Med 2010;29:3245. Pocock et al. Eur Heart J 2012;33:176.*

# GPC – Outcome of any type

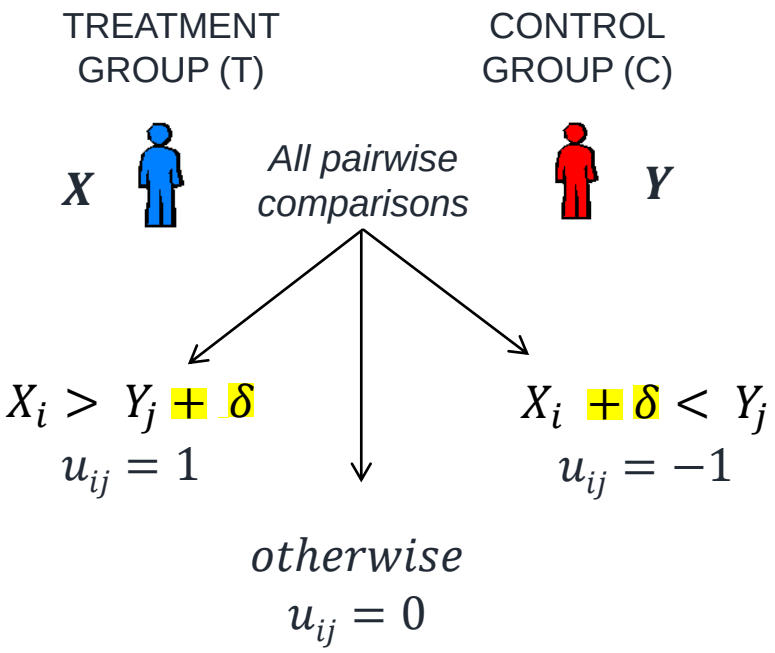



"$X_i$ *better than* $Y_j$" (wins):

- For ordered outcomes, with larger values preferable: $X_i > Y_j$
- For binary outcomes, with 1 denoting success and 0 failure, $X_i > Y_j$
- For time-to-event outcomes, with larger values preferable, $X_i > Y_j$ unless $Y_j$ censored
- For all outcome types, arbitrary definition

*Gehan. Biometrika 1965;52:203*

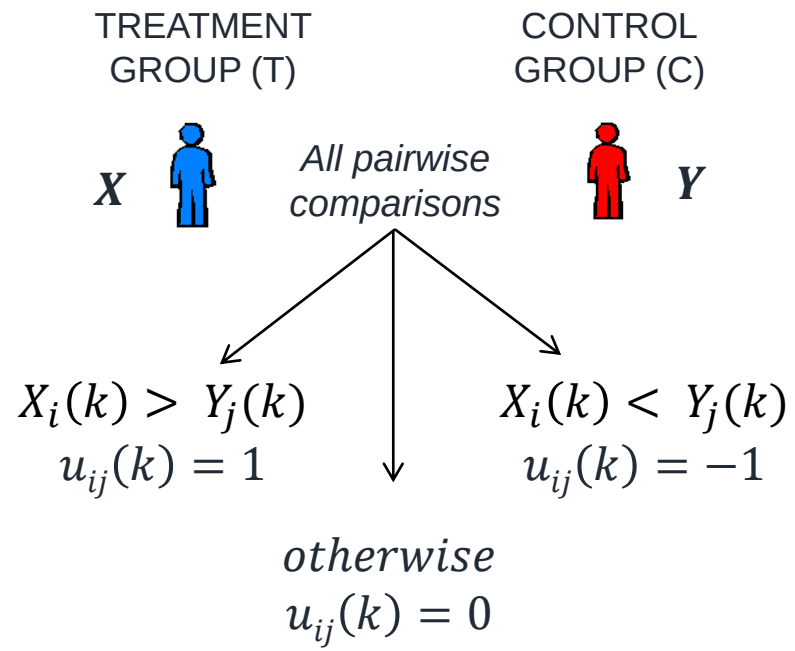
# GPC – clinical threshold




1. Perform pairwise comparisons between all elements of ordered outcomes $\boldsymbol{X}$ and $\boldsymbol{Y}$
2. Calculate $u_{ij} = \begin{cases} +1 \; if \; X_i > Y_j + \delta \\ -1 \; if \; X_i + \delta < Y_j \\ 0 \; otherwise \end{cases}$
3. The statistic
   $U = \frac{1}{m \cdot n} \sum_{i=1}^{m} \sum_{j=1}^{n} u_{ij}$
   has a known distribution under $H_0$

*Buyse. Stat Med 2010;29:3245*

# GPC – multiple weighted outcomes

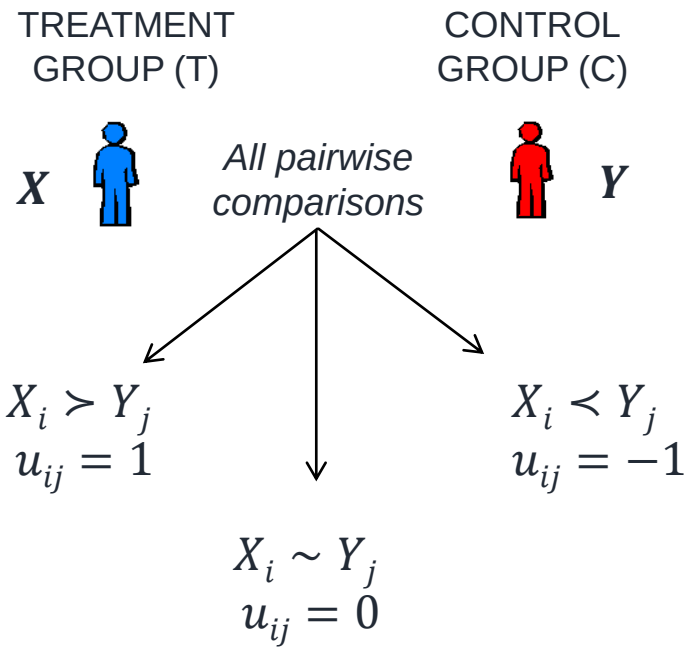



1. Perform pairwise comparisons between all elements of $\boldsymbol{X}$ and $\boldsymbol{Y}$
2. Calculate $u_{ij}(k) = \begin{cases} +1 \; if X_i(k) > Y_j(k) \\ -1 \; if X_i(k) < Y_j(k) \\ 0 \; otherwise \end{cases}$
3. The statistic $U = \frac{1}{m \cdot n} \sum_{k=1}^{K} \sum_{i=1}^{m} \sum_{j=1}^{n} w(k) u_{ij}(k)$ has a known distribution under $H_0$

*Note: weights $w(k)$ are arbitrary, usually chosen so that $\sum_{k=1}^{K} w(k) = 1$*

*O'Brien. Biometrics 1984;69:1079*

# GPC – multiple prioritized outcomes

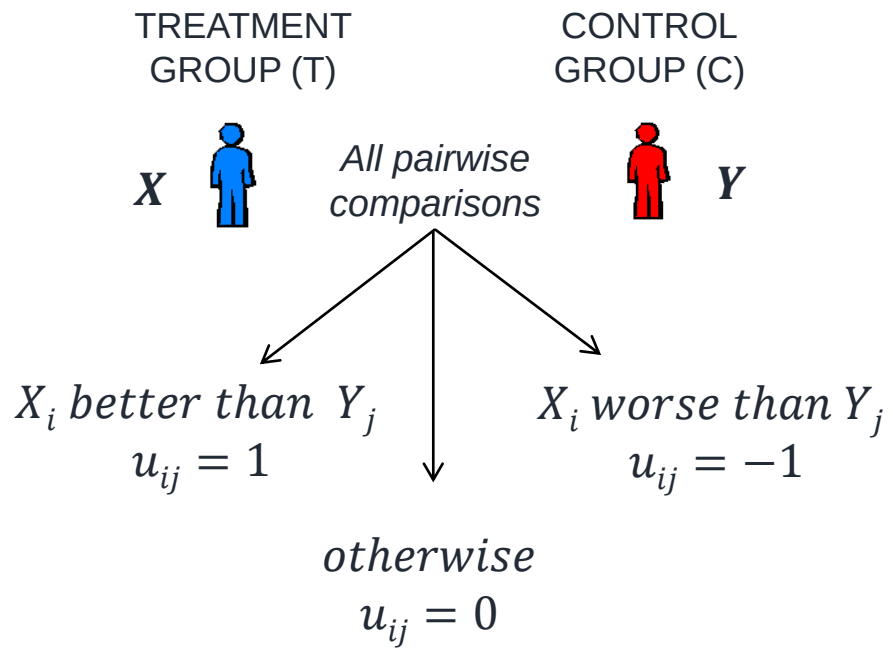



| **Outcome of 1st priority** | **Outcome of 2nd priority** | **Overall** |
|---|---|---|
| Win | - | Win |
| Loss | - | Loss |
| Tie or ? | Win | Win |
| | Loss | Loss |
| | Tie or ? | Tie or ? |

*Note: priorities may be patient-centric*

*Buyse. Stat Med 2010;29:3245*

# Net Treatment Benefit (*NTB*)




The Net Treatment Benefit (*NTB*) is a *U*-statistic

$$U = \frac{1}{m \cdot n} \sum_{i=1}^{m} \sum_{j=1}^{n} u_{ij}$$

$$= \frac{\#Wins - \#Losses}{\#Pairs}$$

*Hoeffding. Ann Math Stat 1948;19:293*

# Measures of treatment effect

$$\textit{Finkelstein-Schoenfeld statistic}^{1} = \#Wins - \#Losses$$

$$NTB^{2} = \frac{\#Wins - \#Losses}{\#Pairs}$$

$$\textit{Win Ratio}^{3} = \frac{\#Wins}{\#Losses}$$

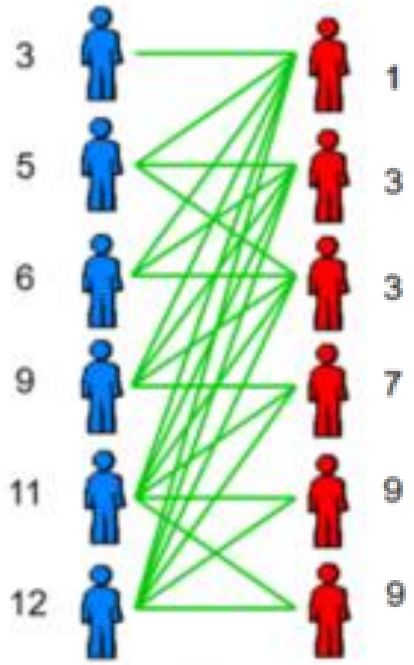
$$\textit{Win Odds}^{4,5} = \frac{\#Wins + \frac{1}{2}\#(Ties\ or\ ?)}{\#Losses + \frac{1}{2}\#(Ties\ or\ ?)}$$

Note

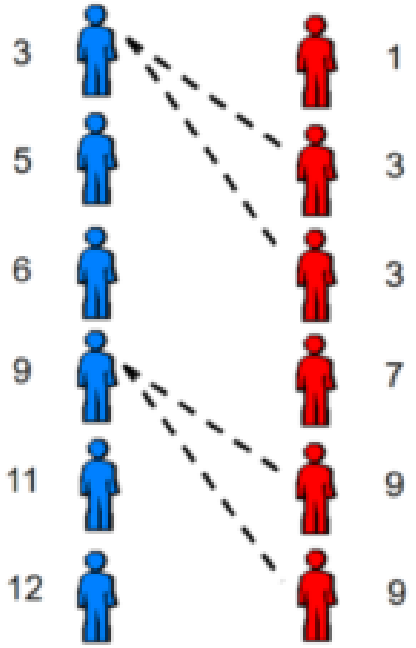
$$NTB = \frac{Win\ Odds - 1}{Win\ Odds + 1}$$

*[1] Finkelstein & Schoenfeld. Stat Med 1999;18:1341. [2] Buyse. Stat Med 2010;29:3245. [3] Pocock et al. Eur Heart J 2012;33:176. [4] Dong et al. Stat Biopharm Res 2020;12:99. [5] Brunner et al. Stat Med 2021;40:3367.*

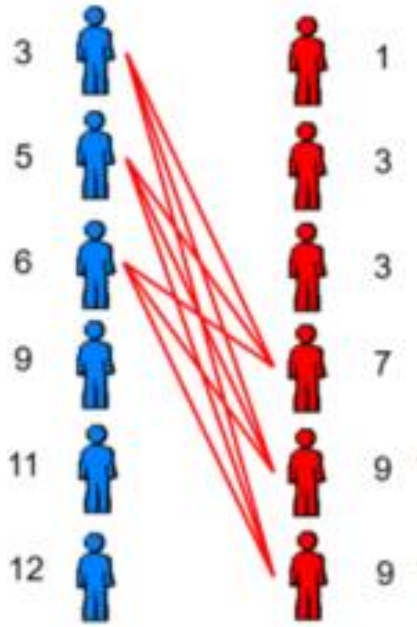
# Measures of treatment effect




Wins




Ties




Losses

$$NTB = \frac{23-9}{36} = 0.39$$

$$Win\ Ratio = \frac{23}{9} = 2.6$$

$$Win\ Odds = \frac{25}{11} = 2.3$$

# $NTB$ – interpretation

$NTB$ ranges from -1 to +1, with 0 indicating no overall treatment effect

$$NTB = P(X > Y) - P(Y > X)$$

$NTB$ is the *net* probability of a better outcome in one treatment group than in the other

More precisely, $NTB$ is the probability that a patient taken at random in the treatment group has a better outcome than a patient taken at random in the control group, minus the probability of the opposite situation.

*Péron et al. J Amer Med Ass 2016;2:905.*

# *NTB* – relationships

$NTB$ is a linear transformation of the probabilistic index $PI$

$$NTB = 2 \cdot PI - 1$$

where

$$PI = P(X > Y) + \frac{1}{2} P(X = Y)$$

$PI$ ranges from 0 to 1, with ½ indicating no overall treatment effect

$PI$ is closely related to the proportion of similar responses [1], the concordance index [2] the probability of overlap [3], and the area under the ROC curve [4].

*[1] Rom & Wang. Stat Med 1996;15:1489. [2] Harrell. Regression Model Strategies, Springer 2001.*
*[3] Stine & Heyse. Stat Med 2001;20:215. [4] Brumback et al. Stat Med 2006;25:575.*

# *NTB* – inference and estimation

For testing $H_0: NTB = 0$, estimation of $NTB$ and confidence limits of $NTB$ [1]:

- Exact permutation and bootstrap distribution of the $NTB$ statistic[2,3]
- Re-randomization tests [4]
- Bootstrapping for confidence intervals [5]
- Asymptotic distribution of $U$-statistics [6-8]

[1] *Verbeeck et al. J Biopharm Stat 2020;30:765.* [2] *Finkelstein & Schoenfeld. Stat Med 1999;18:1341.*
[3] *Anderson & Verbeeck. https://arxiv.org/pdf/1901.10928.pdf, 2019.* [4] *Buyse. Stat Med 2010;29:3245.*
[5] *Pocock et al. Eur Heart J 2012;33:176.* [6] *Dong et al. Pharm Stat 2016;15:430.*
[7] *Bebu & Lachin. Biostatistics 2016;17:178.* [8] *Ramchandani et al. Biometrics 2016;72:926*

# *NTB* – adjustment for censoring

$NTB$ (Gehan Wilcoxon test) is biased in the presence of censoring [1]. The bias can be removed through different approaches [2]

- Naïve, using the proportion of informative pairs [3,4]
- Imputations using the survival distribution [1,5,6]
- Inverse probability of censoring weighting [7,8]

*[1] Efron. Proc 5th Berkeley Symp 1967;4:831. [2] Deltuvaite-Thomas et al. Biometrical J 2022.*
*[3] Harrell et al. J Am Med Ass 1982;247:2543. [4] Buyse. Clin Trials 2008;5:641.*
*[5] Latta. Biometrika 1977;63:633. [6] Péron et al. Stat Meth Med Res 2016;27:1230.*
*[7] Datta et al. Scand J Stat 2010;37:680. [8] Dong et al. Stat Biopharm Res 2020;30:882*

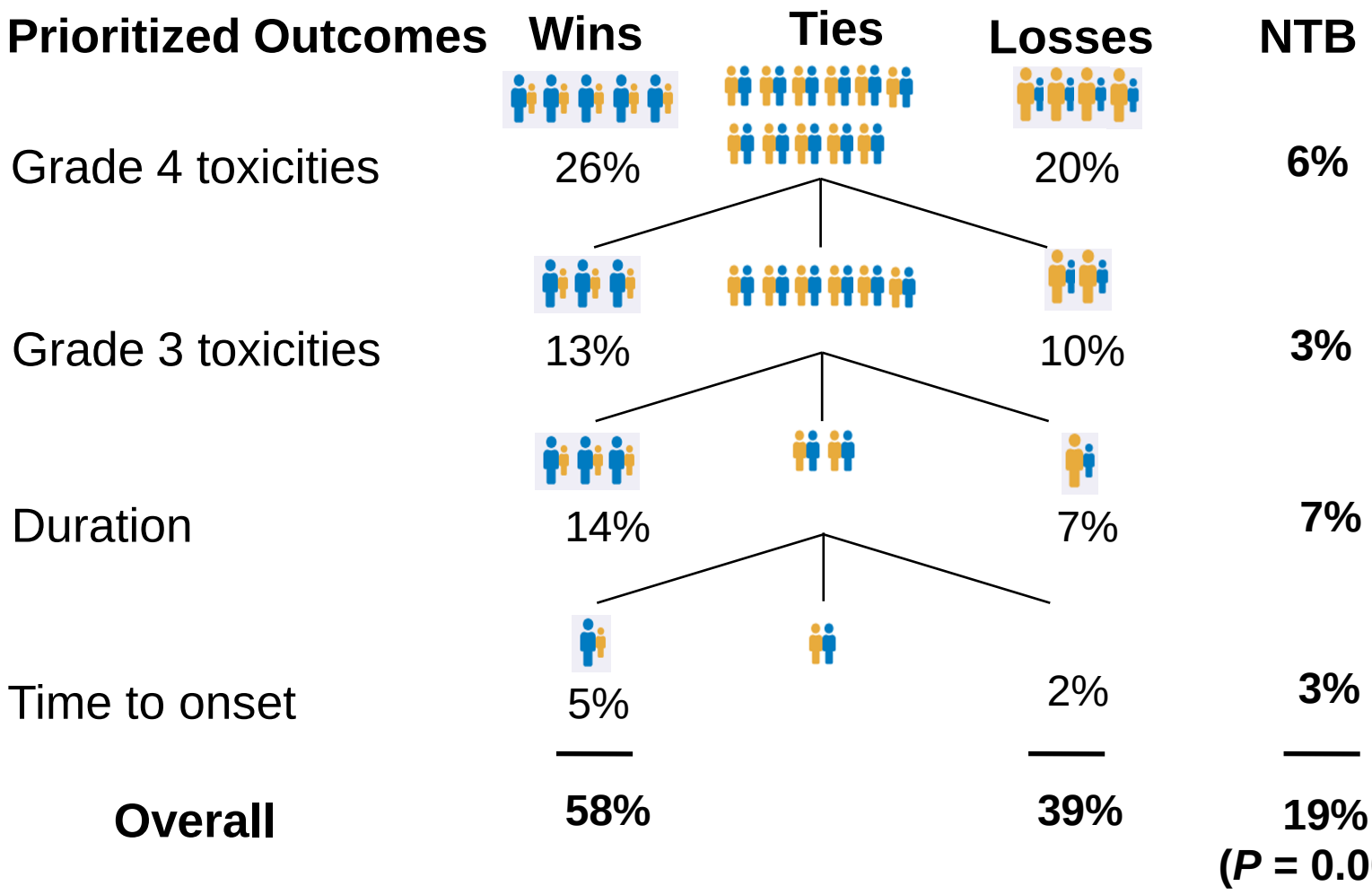
# Applications

# Augmenting power *and* clinical relevance

- Patients with cancer treated aggressively may experience severe toxicities
  - WHO grade 3: severe
  - WHO grade 4: life-threatening
  - WHO grade 5: lethal
- The traditional primary endpoint for comparing an experimental treatment with a control is incidence of WHO grade 3 or worse toxicity
- The analysis should take multiple prioritized outcomes into account:
  1. Severity (lower WHO grade better)
  2. Duration of severe toxicity (shorter better)
  3. Time to onset (later better)

# Augmenting power *and* clinical relevance

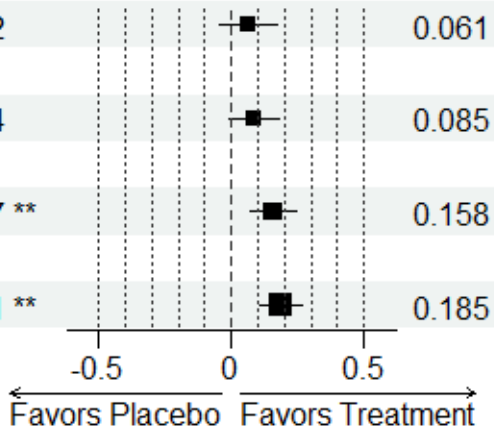
Placebo controlled trial of experimental treatment protecting against a specific toxicity



# Augmenting power *and* clinical relevance

Placebo controlled trial of experimental treatment protecting against a specific toxicity

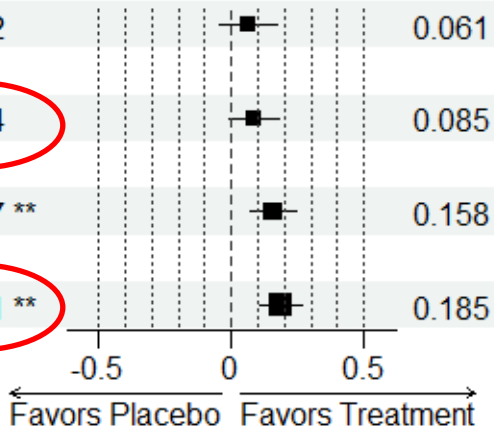
| Outcome | Pairs | Wins(%) | Losses(%) | Ties(%) | P-value | NTB (95% CI) |
|---|---|---|---|---|---|---|
| Grade 4 toxicities | 1000 | 26.5 | 20.4 | 53.1 | 0.072 | 0.061 (-0.049 to 0.171) |
| Grade 3 or 4 toxicities | 531 | 12.6 | 10.2 | 30.3 | 0.054 | 0.085 (-0.006 to 0.176) |
| Duration | 303 | 14.1 | 6.8 | 9.4 | 0.007 ** | 0.158 (0.073 to 0.243) |
| Time to Onset | 94 | 4.5 | 1.8 | 3.1 | 0.001 ** | 0.185 (0.104 to 0.266) |



# Augmenting power *and* clinical relevance

Placebo controlled trial of experimental treatment protecting against a specific toxicity

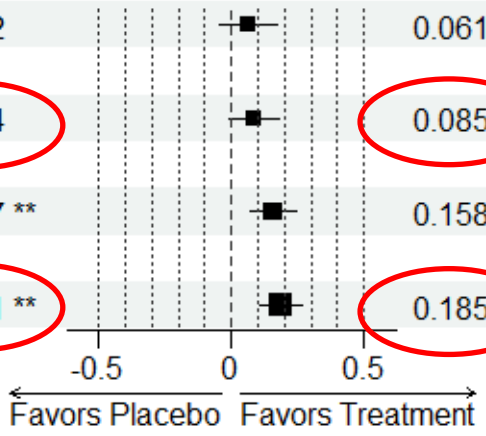
| Outcome | Pairs | Wins(%) | Losses(%) | Ties(%) | P-value | NTB (95% CI) |
|---|---|---|---|---|---|---|
| Grade 4 toxicities | 1000 | 26.5 | 20.4 | 53.1 | 0.072 | 0.061 (-0.049 to 0.171) |
| Grade 3 or 4 toxicities | 531 | 12.6 | 10.2 | 30.3 | 0.054 | 0.085 (-0.006 to 0.176) |
| Duration | 303 | 14.1 | 6.8 | 9.4 | 0.007 ** | 0.158 (0.073 to 0.243) |
| Time to Onset | 94 | 4.5 | 1.8 | 3.1 | 0.001 ** | 0.185 (0.104 to 0.266) |



# Augmenting power *and* clinical relevance

Placebo controlled trial of experimental treatment protecting against a specific toxicity

| Outcome | Pairs | Wins(%) | Losses(%) | Ties(%) | P-value | NTB (95% CI) |
|---|---|---|---|---|---|---|
| Grade 4 toxicities | 1000 | 26.5 | 20.4 | 53.1 | 0.072 | 0.061 (-0.049 to 0.171) |
| Grade 3 or 4 toxicities | 531 | 12.6 | 10.2 | 30.3 | 0.054 | 0.085 (-0.006 to 0.176) |
| Duration | 303 | 14.1 | 6.8 | 9.4 | 0.007 ** | 0.158 (0.073 to 0.243) |
| Time to Onset | 94 | 4.5 | 1.8 | 3.1 | 0.001 ** | 0.185 (0.104 to 0.266) |



# Benefit / risk analysis

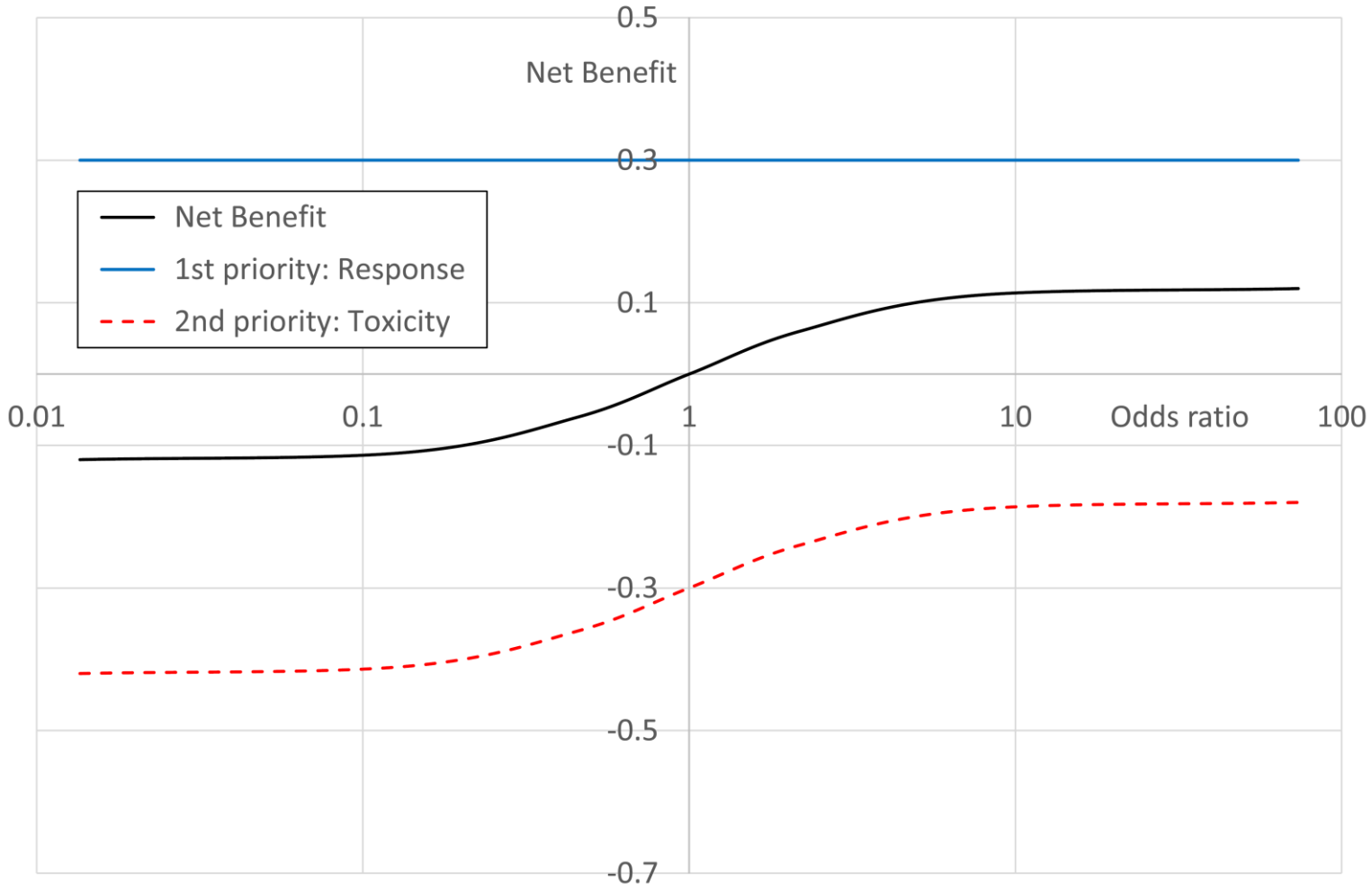
- Simple situation of binary efficacy outcome (1 = response, 0 = no response) and binary safety outcome (1 = no toxicity, 0 = toxicity)

| **Outcomes** | **Treatment** | **Control** | **Difference** |
|---|---|---|---|
| Response rate (benefit) | 0.5 | 0.2 | 0.3 |
| Toxicity rate (risk) | 0.6 | 0 | 0.6 |
| Marginal benefit / risk difference | | | -0.3 |

- Naïve analysis suggests negative benefit / risk of -0.3
- What would GPC analysis show, assuming achievement of response is prefered to avoidance of toxicity?

*Buyse et al. J Clin Epidemiol 2021;137:148*

# Benefit / risk analysis




*Buyse et al. J Clin Epidemiol 2021;137:148*

# Benefit / risk analysis

- $NTB$ depends on the association (odds ratio, $OR$) between response and toxicity
  - If $OR > 1, NTB > 0$ : patients who respond also have toxicity (*e.g.*, skin rash for inhibitors of the EGFR pathway)
  - If $OR = 1, NTB = 0$ : response is independent of toxicity (*e.g.*, cardiac toxicities of anthracyclins)
  - If $OR < 1, NTB < 0$ : patients who do not respond have toxicity (*e.g.*, toxicities to irinotecan in patients with enzyme deficiencies)

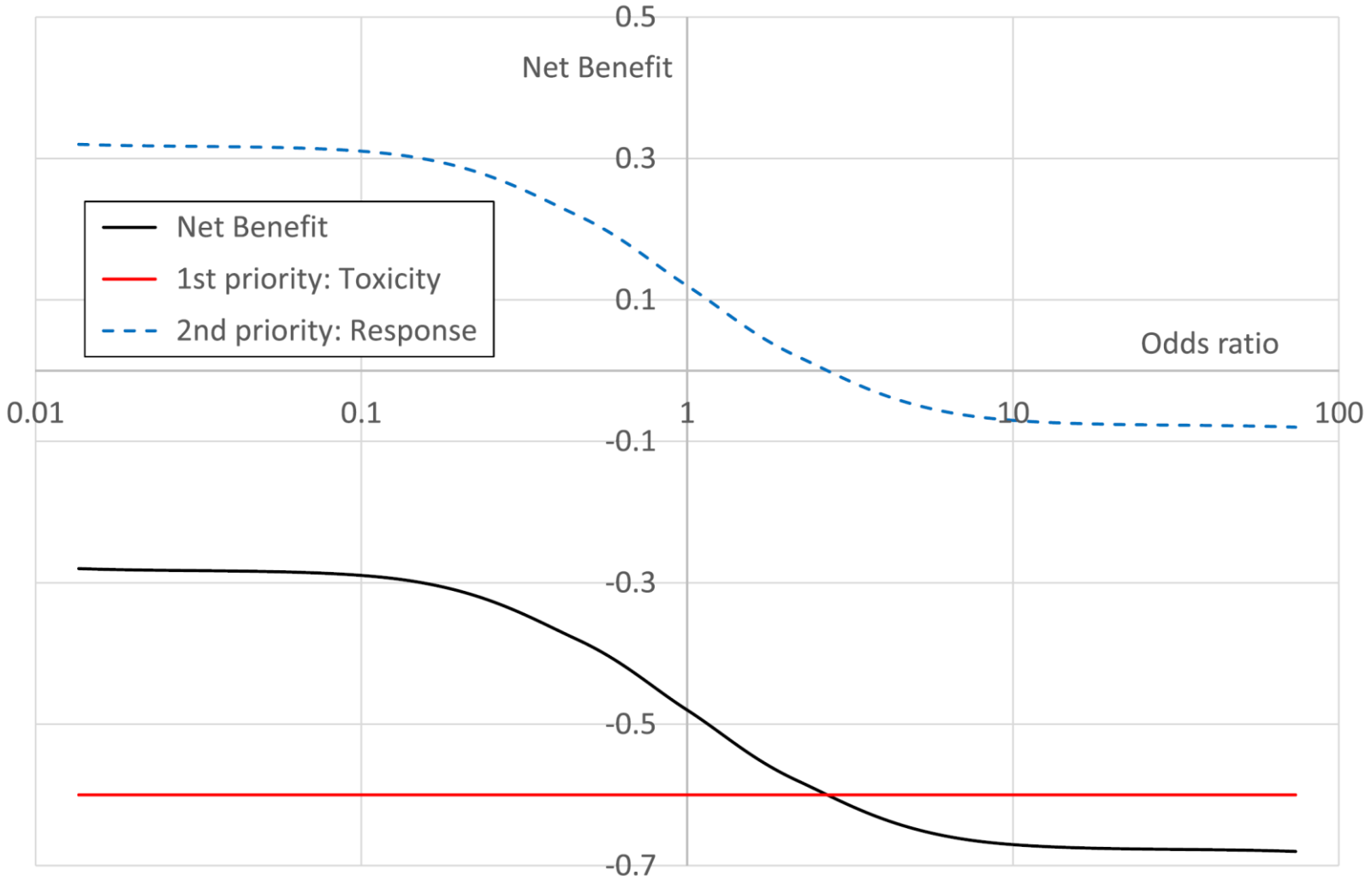
*Buyse et al. J Clin Epidemiol 2021;137:148*

# Benefit / risk analysis

- $NTB$ depends on the association (odds ratio, $OR$) between response and toxicity
  - If $OR > 1, NTB > 0$ : patients who respond also have toxicity (*e.g.*, skin rash for inhibitors of the EGFR pathway)
  - If $OR = 1, NTB = 0$ : response is independent of toxicity (*e.g.*, cardiac toxicities of anthracyclins)
  - If $OR < 1, NTB < 0$ : patients who do not respond have toxicity (*e.g.*, toxicities to irinotecan in patients enzyme deficiencies)
- $NTB$ would be quite different if avoidance of toxicity was prefered to achievement of response, allowing for patient-centric treatment choices

*Buyse et al. J Clin Epidemiol 2021;137:148*

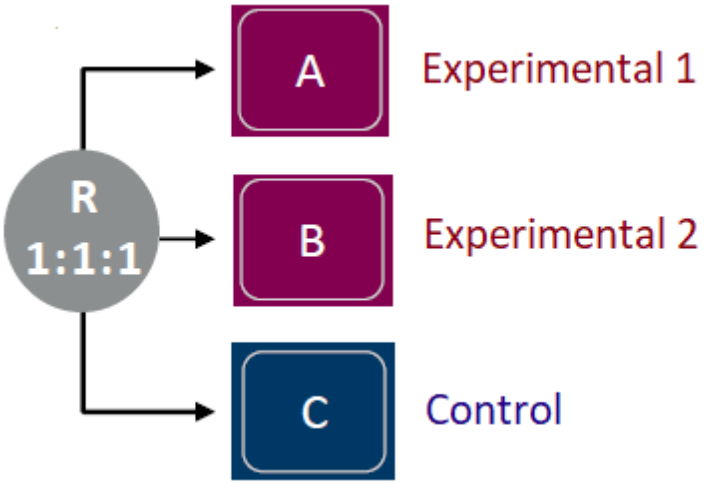
# Benefit / risk analysis




*Buyse et al. J Clin Epidemiol 2021;137:148*

# Multiple Testing Procedures

Assume several treatments are compared to a standard of care

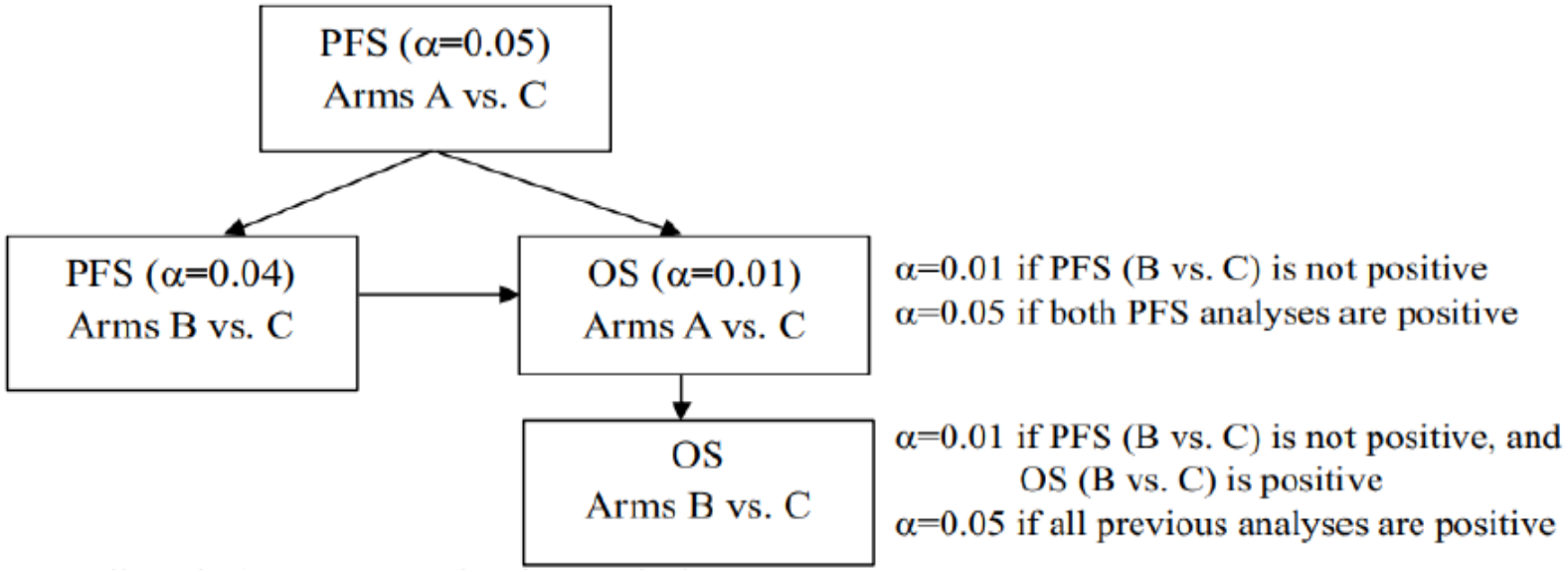



Comparisons: A vs. C (Experimental 1, preferred)
B vs. C (Experimental 2)
A vs. B (Not powered)

Outcomes: PFS (« Primary »)
OS (« Key secondary »)

# Multiple Testing Procedures

Testing procedure with strict control of type I error rate



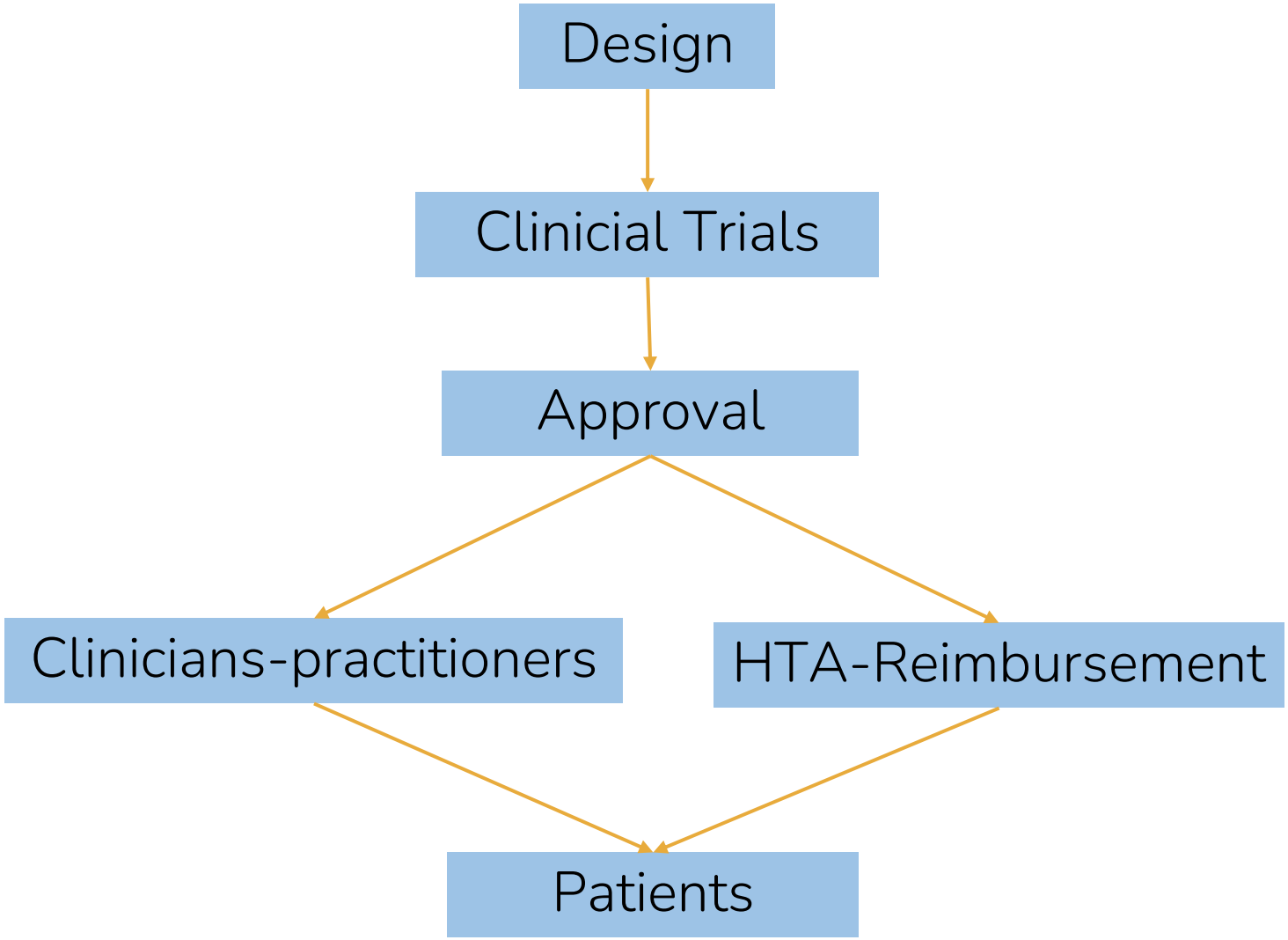

OS of the preferred experimental arm is tested at full level of significance (0.05) *only if PFS of the other (non preferred) experimental arm reaches statistical significance !*

# Conclusions

# GPC benefits

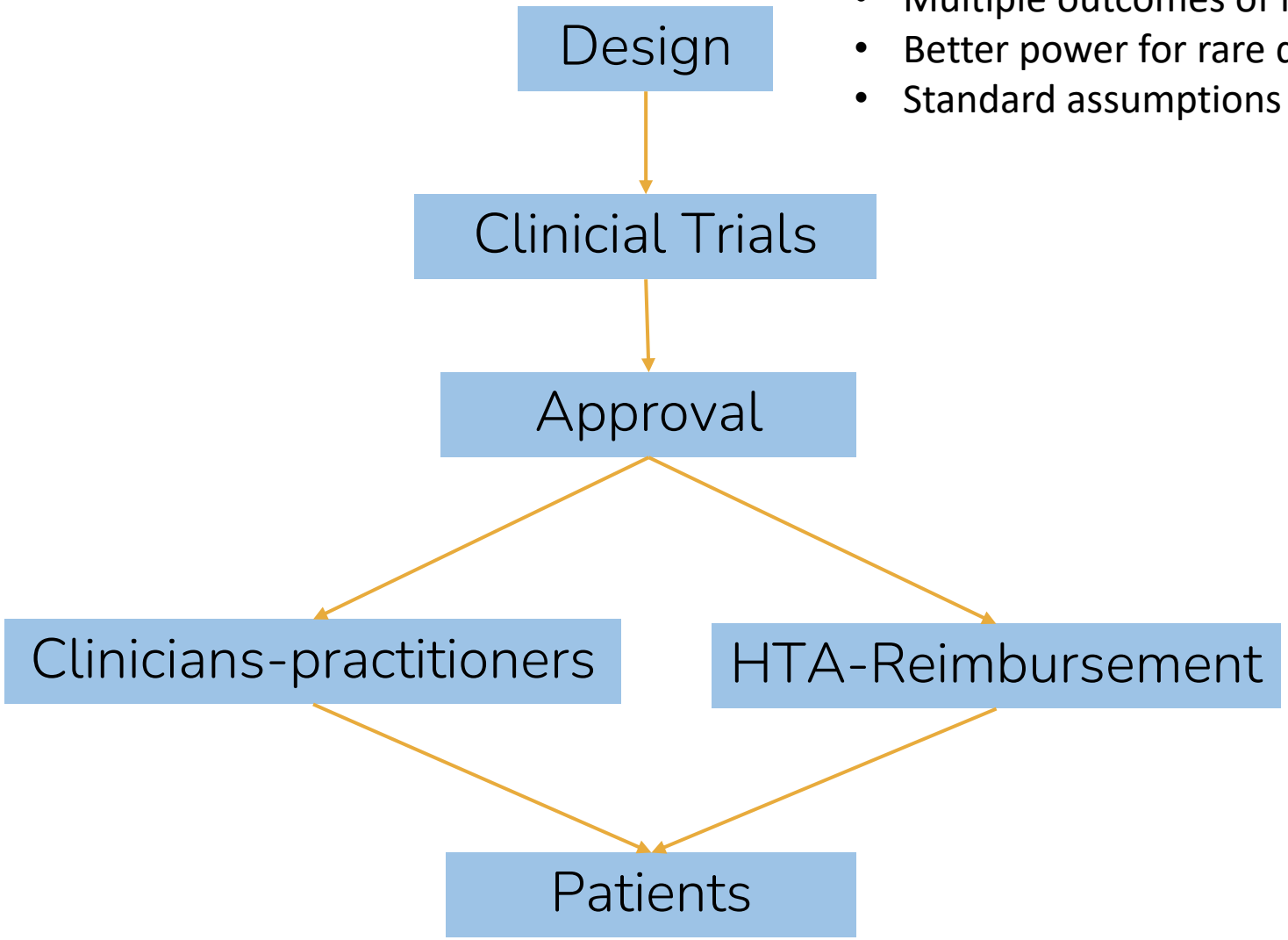
- Increases flexibility of analyses
- Incorporates multiple outcomes
- Incorporates thresholds of clinical relevance
- May increase power as compared with single outcome
- Can be adapted to individual patient preferences
- Provides unique measure of treatment effect that is meaningful to patients and caregivers

# GPC usage


Design
Clinicial Trials
Approval
Clinicians-practitioners
HTA-Reimbursement
Patients

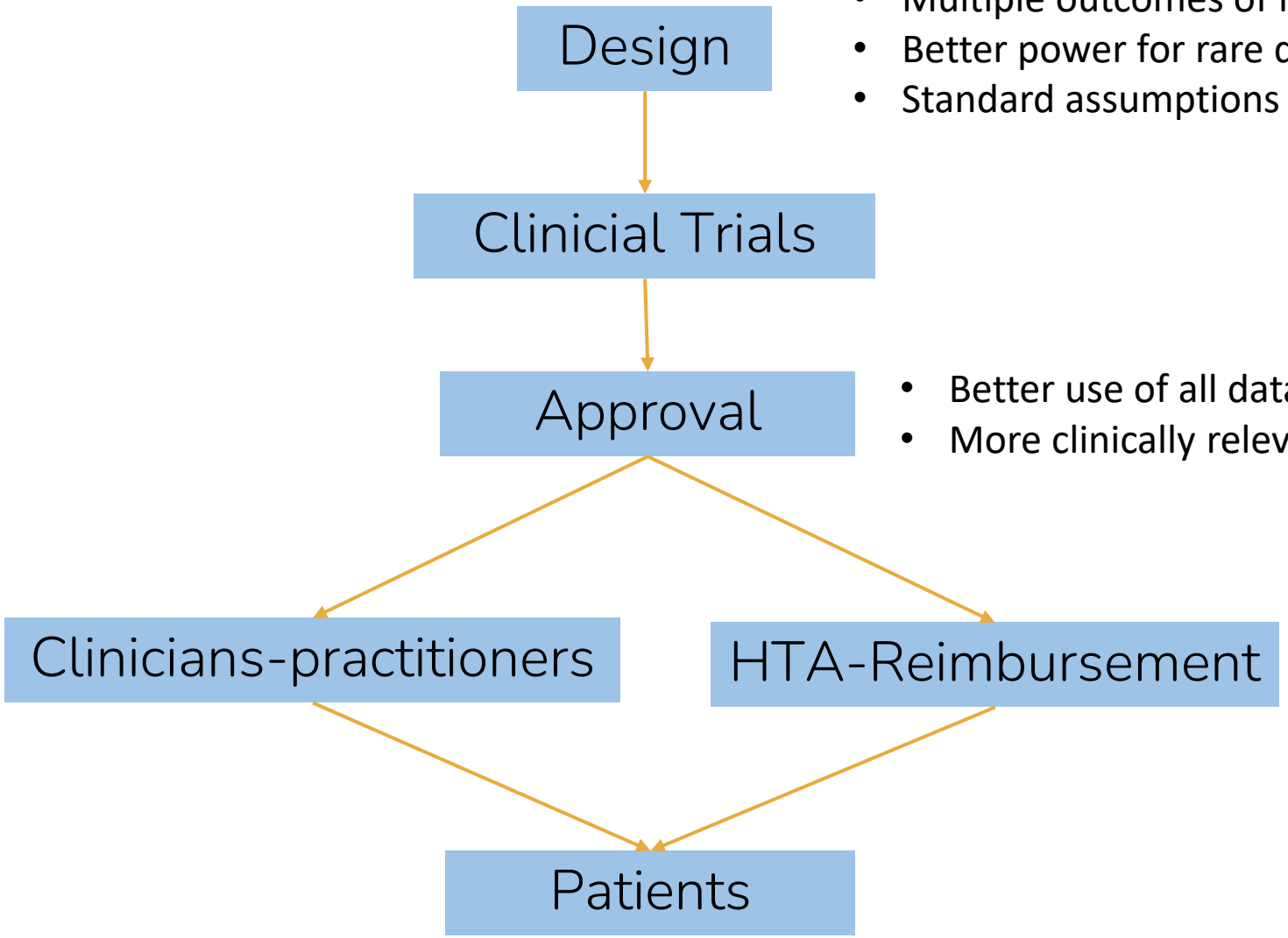
# GPC usage

- Multiple outcomes of interest
- Better power for rare diseases
- Standard assumptions not met



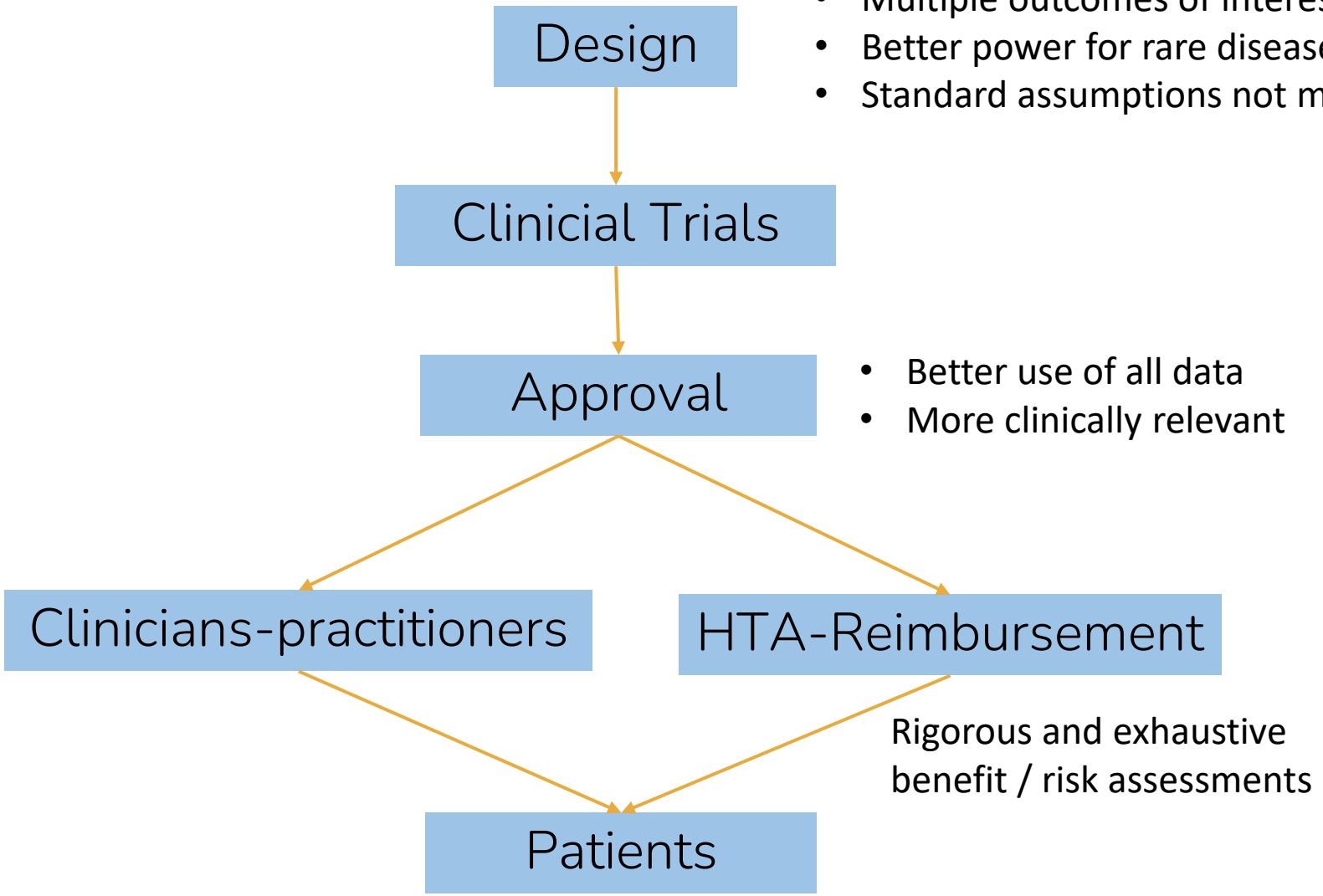
# GPC usage

- Multiple outcomes of interest
- Better power for rare diseases
- Standard assumptions not met




- Better use of all data
- More clinically relevant

# GPC usage

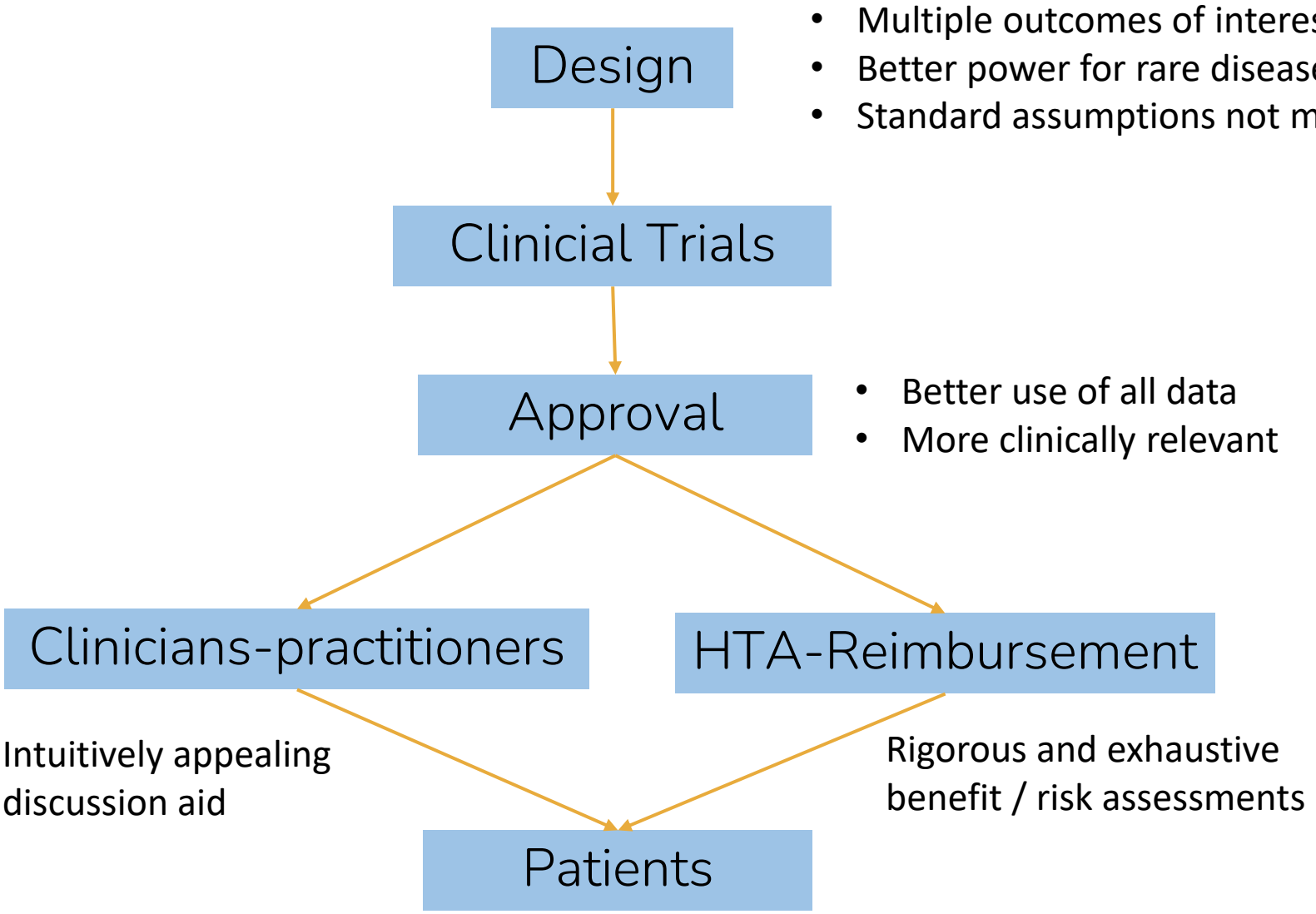



- Multiple outcomes of interest
- Better power for rare diseases
- Standard assumptions not met

- Better use of all data
- More clinically relevant

Rigorous and exhaustive benefit / risk assessments

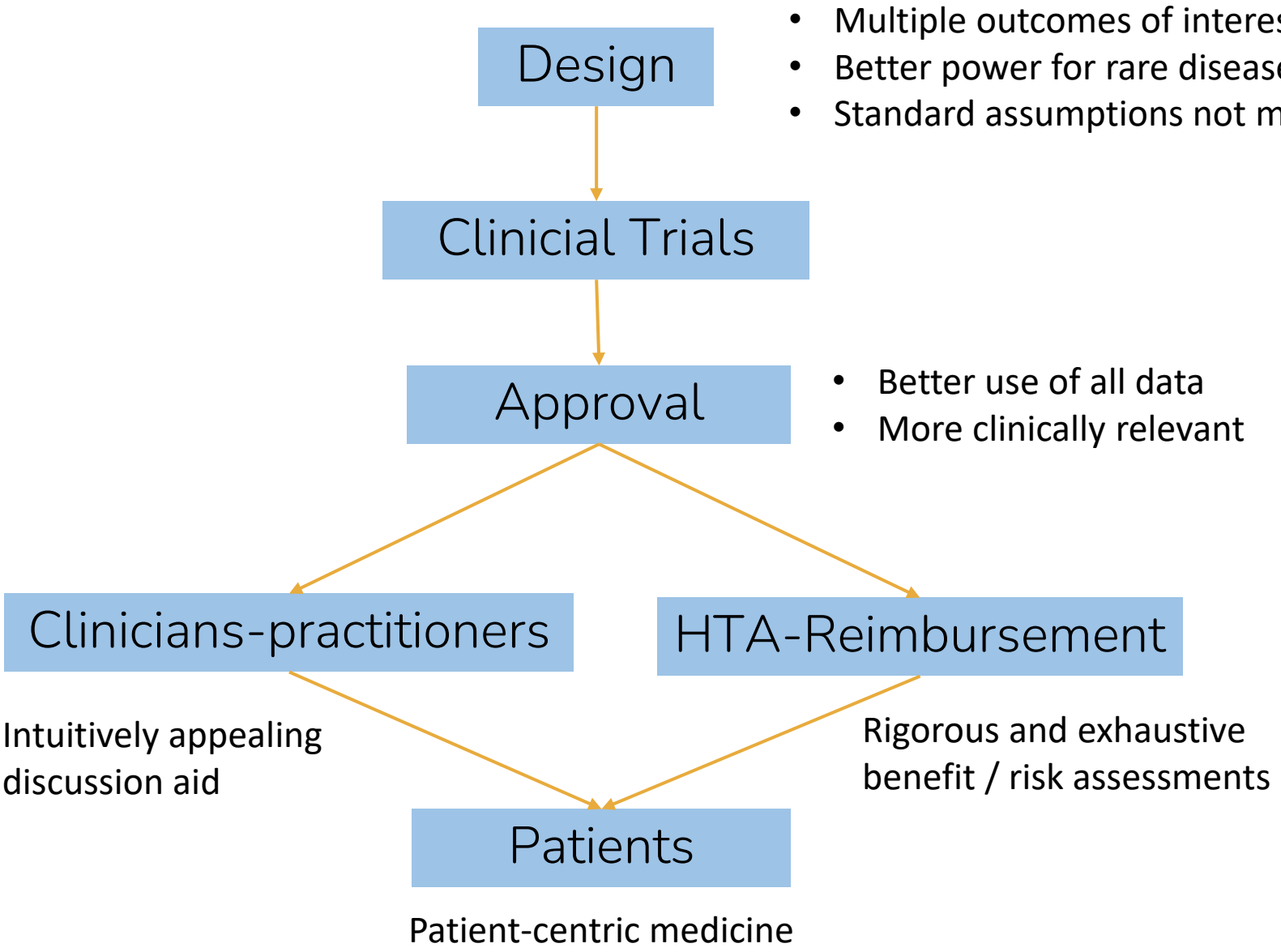
# GPC usage




- Multiple outcomes of interest
- Better power for rare diseases
- Standard assumptions not met

- Better use of all data
- More clinically relevant

# GPC usage




- Multiple outcomes of interest
- Better power for rare diseases
- Standard assumptions not met

- Better use of all data
- More clinically relevant

# Questions / References

marc.buyse@iddi.com