THE USE OF GENERALIZED PAIRWISE COMPARISONS FOR DESIGNING TRIALS THAT ARE TAILORED TO PATIENTS' WISHES AND NEEDS

Mickaël De Backer

20. September 2022

IDDI, International Drug Development Institute

Overview

Structure

1. Motivating examples
2. Non-inferiority criticism
3. Generalized Pairwise Comparisons
4. Application

1. MOTIVATING EXAMPLES

1. Motivating examples

Example 1

Context

- Disease: locally advanced rectal cancer
- Population of interest: elderly ($\geq$ 70 years)
- Standard of care: short course radiotherapy + cycles of chemotherapy

Motivation

- Elderly population is more fragile
- The benefits from the additional chemotherapy may be overthrown by its harms in toxicity
- New suggested treatment: only radiotherapy

  $\rightarrow$ Context for a **non-inferiority trial**?

Example 1

Assumptions of the non-inferiority trial

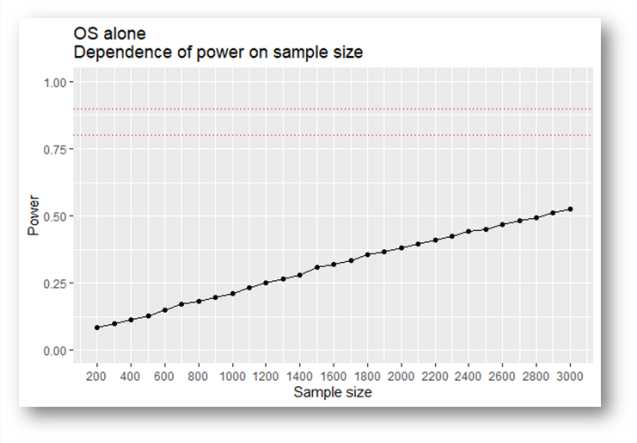
$$\begin{cases} \text{Trt: Short course radiotherapy (SCRT)} \\ \text{Ctr: Rapido - SCRT + 6 capox} \end{cases}$$

- Potential endpoints of interest:
  - T_1: overall survival at 36 months
  - T_2: disease recurrence at 36 months
- Proportions of events at 36 months:

	Trt	Ctr
T_1	15%	11%
T_2	32%	24%

Example 1 - Questions

1. Is the trial feasible?



2. How to include patients' voices in the design?

Example 2

Context

- Disease: low and intermediate risk acute promyelocytic leukemia
- Standard of care: full dose of ATRA (all-trans retinoic acid)

Motivation

- Full dose induces much toxicities
- Reported real world data shows that a reduced dose may provide similar efficacy
- New suggested treatment: lowered dose

  → Context for a **non-inferiority trial**?

Example 2

Assumptions of the non-inferiority trial

$$\begin{cases} \text{Trt: Reduced ATRA dose} \\ \text{Ctr: Full ATRA dose} \end{cases}$$

- Potential endpoints of interest:
  - T_1: event-free survival at 24 months
- Proportions of events at 24 months:

	Trt	Ctr
T_1	10%	8%

Example 2 - Questions

1. Is the trial feasible?

Outcome	Value in Control Arm*	Value in Experimental Arm**	Difference (Control – Experimental)	Non-inferiority margin	Approximate Sample Size
EFS at 2 years non-inferior	0.92	0.92	0.0	0.05	1000
	0.92	0.91	0.01	0.05	1500
	0.92	0.90	0.02	0.05	2900

2. How to include patients' voices in the design?

2. NON-INFERIORITY CRITICISM

2. Non-inferiority criticism

Context

- Our trials are infeasible in terms of sample size and leave no room to the voice of patients.
- But there are other regular criticism towards non-inferiority trials:
  - Composite objective (acceptable deterioration is compensated), yet benefits are either (i) not shown, or (ii) analyzed separately with disregard of dependencies

Toxicity	
Absence	Presence
50%	50%

Survival	
Do worse	Do better
50%	50%

?

?

		Survival	
		Do worse	Do better
Toxicity	Absence	50%	0%
	Presence	0%	50%

		Survival	
		Do worse	Do better
Toxicity	Absence	0%	50%
	Presence	50%	0%

2. Non-inferiority criticism

Context

- There are other regular criticism towards non-inferiority (NI) trials:
  - Assay sensitivity (capacity of a trial to distinguish an effective therapy from one that is not effective); NI trials are biased towards positive results if they are poorly designed and conducted.
    Review study of Wangge et al. (2010): 1/3 of the 232 PubMed registered trials did not use blinding.
  - Choice of population: ITT may bias towards positive result, while per protocol has smaller size
  - Choice of the margin: no statistical consensus, prone to 'cherry-picking', and not easy to include patients' voices
  - 'biocreep' phenomenon

$\Rightarrow$ Can we address these concerns with a statistically sound procedure?

3. GENERALIZED PAIRWISE COMPARISONS

Context and Motivation

Consider a randomized clinical trial with a treatment and a control group:

- Let Y^t be the continuous outcome of interest in the treatment group.
- Let Y^c be the continuous outcome of interest in the control group.
- We observe Y_i^ℓ, $i = 1, \ldots, n^\ell$, $\ell \in \{t, c\}$.
- Question of the trial: is there a treatment effect?
- One way to assess: the Mann-Whitney form of the Wilcoxon test

$$U_{ij} = \begin{cases} 1 & \text{if } Y_i^t > Y_j^c \\ 1/2 & \text{if } Y_i^t = Y_j^c \\ 0 & \text{if } Y_i^t < Y_j^c \end{cases}$$

Test is then based on

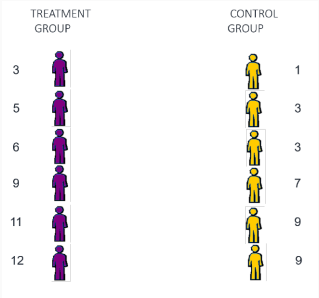
$$U = \frac{1}{n^t n^c} \sum_{i=1}^{n^t} \sum_{j=1}^{n^c} U_{ij}.$$

Context

$$U = \frac{1}{n^t n^c} \sum_{i=1}^{n^t} \sum_{j=1}^{n^c} U_{ij}.$$

Way of reading:

- Simple idea of comparing pairs of patients: win, loss, neutral
- Besides test, effect size measured is $\mathbb{P}(Y^t > Y^c) + 1/2\,\mathbb{P}(Y^t = Y^c)$



Idea

Generalize the pairwise comparisons:

- 'Classify' for all ordinal types of outcomes, e.g.
  - Binary: patient alive is better than dead.
  - Continuous: e.g. win if outlive for at least τ months, $Y_i^t \geq Y_j^c + \tau$.
- Include multiple outcomes to keep classifying pairs:

First priority outcome	Second priority outcome	Classification
Win	-	Win
Loss	-	Loss
Tie/uninformative	Win	Win
Tie/uninformative	Loss	Loss
Tie/uninformative	Tie/uninformative	Tie/uninformative

The Net Treatment Benefit

- Let Y_i^t and Y_j^c be possibly multivariate:

$$U_{ij} = \begin{cases} +1 & \text{if } Y_i^t \succ Y_j^c \\ -1 & \text{if } Y_i^t \prec Y_j^c \\ 0 & \text{otherwise.} \end{cases}$$

  where $\succ$ and $\prec$ define wins and losses according to the context.

- Generalized Pairwise Comparisons:

$$\widehat{\Delta}_\tau := \frac{1}{n^t n^c} \sum_{i=1}^{n^t} \sum_{j=1}^{n^c} U_{ij},$$

  is an estimator of $\Delta_\tau := \mathbb{P}(Y^t \succ Y^c) - \mathbb{P}(Y^t \prec Y^c)$, called the Net Treatment Benefit.

- Other GPC statistics:

  Win Ratio $:= \mathbb{P}(Y^t \succ Y^c) \,/\, \mathbb{P}(Y^t \prec Y^c)$
  Win Odds $:= \{\mathbb{P}(Y^t \succ Y^c) + 1/2\mathbb{P}(Y^t \asymp Y^c)\} \,/\, \{\mathbb{P}(Y^t \prec Y^c) + 1/2\mathbb{P}(Y^t \asymp Y^c)\}$

Theoretical Developments

GPC-based statistics have known many theoretical developments that make them now available for designing practical trials. For instance (non-exhaustive list):

- Adaptations to nuisances / particularities in the data:
  - Censoring: Gehan (1965), Efron (1967), Péron et al. (2016), Deltuvaite-Thomas et al. (2022)
  - Missing Data: Deltuvaite-Thomas and Burzykowski (2022)
  - Competing Risks: Cantagallo et al. (2020)
- Inferential procedures:
  - Asymptotic procedures: Bebu and Lachin (2016), Ozenne et al. (2022)
  - Resampling procedures: Buyse (2010), Pocock et al. (2012), Chung and Romano (2016), Anderson and Verbeeck (2019)

4. APPLICATION

Example 1

- Endpoints of interest:
  - T_1: Survival, time to death from any cause.
  - T_2: Time to disease recurrence.
  - X_1: Any grade ≥ 2 neurological toxicity ($0 \succ 1$).
  - X_2: Any grade ≥ 3 toxicities during pre-operative treatment ($0 \succ 1$).

Natural prioritization: efficacy $\succ$ toxicities

- Incorporating the voice of patients:
  - Prioritization of interest:

$$T_2 \succ T_1 \succ X_1 \succ X_2$$

  - For toxicities to be ignored, need a efficacy gains of at least 12 months
  - Final prioritization:

$$T_2[12] \succ T_1[12] \succ X_1 \succ X_2$$

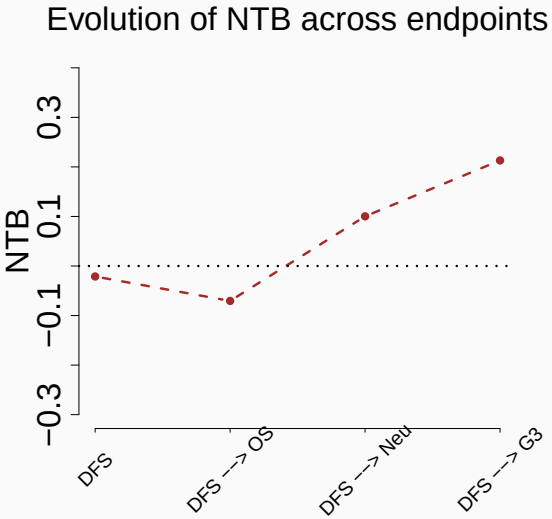
Example 1 - assumptions

1. Proportions of events at 36 months:

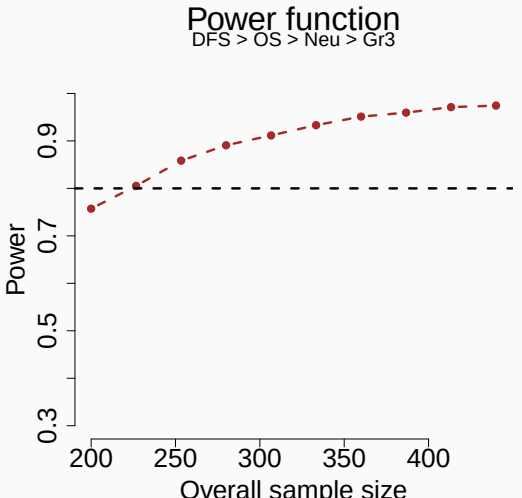
	Trt	Ctr
T_1: death at 36 months	15%	11%
T_2: disease recurrence at 36 months	32%	24%
X_1: $P(X_1 = 1)$ (lower better)	5%	35%
X_2: $P(X_2 = 1)$ (lower better)	24%	55%

2. $T_1, T_2 \sim Exp(\cdot)$
3. $X_1, X_2 \sim Bernoulli(\cdot)$
4. Same dependence structure for the two groups

Example 1 - Net Treatment Benefit

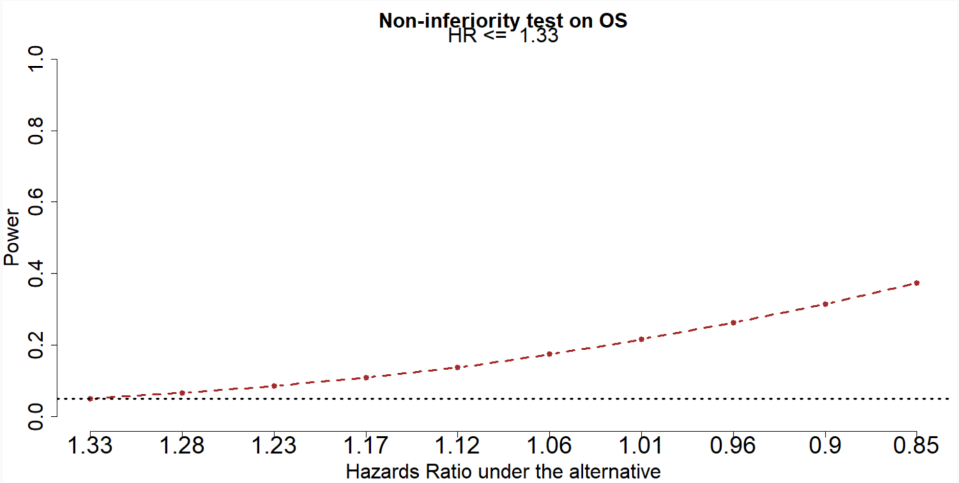


Example 1 - Sample size



Example 1 - Comparing with a conventional NI trial

Using the sample size obtained by the GPC-based design:



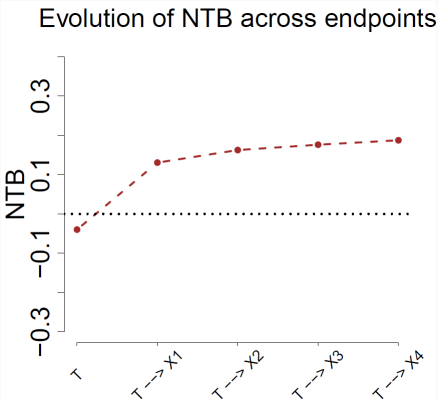
Example 1 - Lessons

- Non-inferiority concerns:
  - ✓ sample size
  - ✓ voice of patients
  - ✓ composite objective
  - ✓ population of interest
  - ✓ choice of margin
  - ✓ biocreep
  - assay sensitivity: can we prevent bias towards positive results for sloppy trials?

Example 2

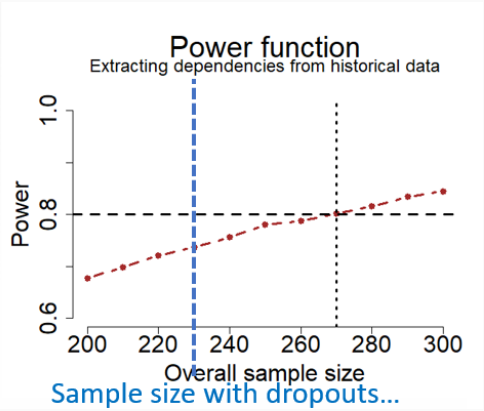
- Endpoints of interest:
  - T_1: EFS, binary ($0 \succ 1$).
  - X_1 till X_4: toxicities ($0 \succ 1$).

Natural prioritization: $T_1 \succ$ toxicities



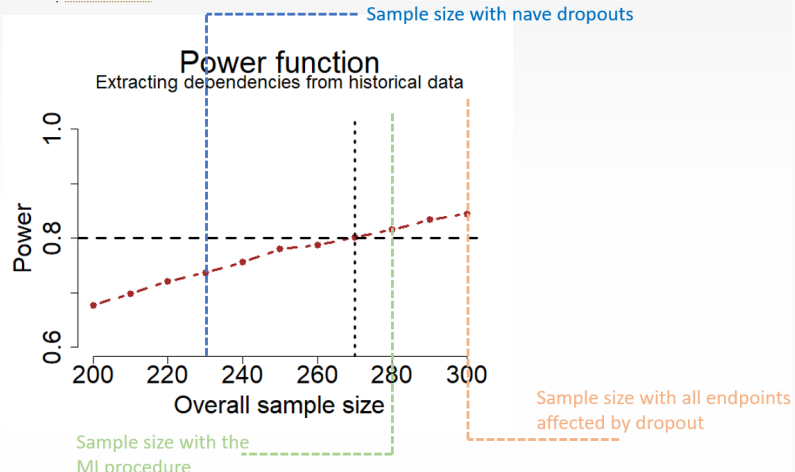
Example 2

- The impact of time:
  - Toxicities (in favor of the treatment) are measured early on
  - EFS (in favor of the control) is measured at end of trial
  - If naive handling of drop-outs:



Example 2

- The impact of time:
  - We may enforce the penalization of sloppy data collection, hence avoiding assay sensitivity issues:



5. CONCLUSION

Conclusion

- Generalized Pairwise Comparisons provide a simple tool for designing trials with:
  - realistic sample sizes
  - explicit composite objective
  - incorporation of the voice of patients
- Recent theoretical developments provide the framework for these designs in many situations, but challenges remain:
  - modelling of dependence structures
  - consideration of more complex designs (interim analysis etc)

Thanks!
Questions?