Multiplicity Control in Oncology Clinical Trials with a Surrogate Endpoint-Based Drop-the-Losers Design

ASA NJ and Princeton Chapter Spring Symposium

CG Wang
Head of Statistical Innovation
Regeneron Pharmaceuticals, Inc.

Introduction

Phase I and II Oncology Studies

- In traditional oncology drug development, dose-finding is primarily conducted in phase 1 clinical studies
  - The maximum tolerated dose is identified as the "optimal" dose
- However, with the increasing use of new molecular targeted agents and immunotherapies, the importance of dose-finding trials has become paramount, as highlighted in the *Project Optimus*
- Consequently, phase 2 oncology clinical trials has been expanded to include the selection of "optimal" doses
  - Multiple doses, randomized

Existing Approaches

- *Multiple test-based procedures* (Budde and Bauer 1989; Bretz, Pinheiro, and Branson 2005; Wakana, Yoshimura, and Hamada 2007)
- *Adaptive seamless multi-arm multi-stage trials* (Posch et al. 2005; Franchetti, Anderson, and Sampson 2013; Schmidli, Bretz, and Racine-Poon 2007; Pozzi et al. 2013; J. M. Wason and Trippa 2014)
  - *Drop-the-losers design* (Sampson and Sill 2005; Sill and Sampson 2009; J. Wason et al. 2017; Abbas et al. 2022)
  - Multiple treatment arms are administered during the initial stage. The most effective treatment arm, as determined by *pre-specified selection criteria*, is then chosen for further stages

Research Question

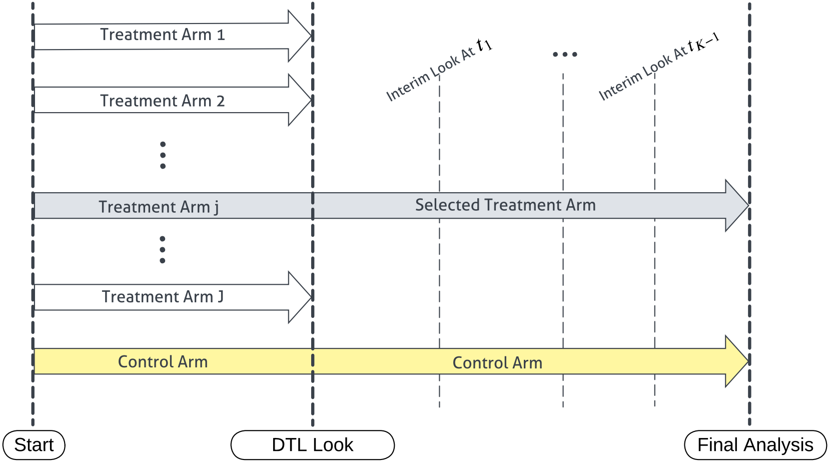
- How to design an oncology trial with DTL that is based on binary surrogate endpoint, such as overall response (OR) or pathological complement response (CR)?
- Outline
    - Evaluate the “correlation” between binary surrogate endpoints and PFS or OS
    - Examine the impact of this correlation on type I error inflation
    - Derive and validate a theoretical boundary for this correlation
    - Assess the performance of the proposal in various scenarios

Method

Drop-The-Losers Design i

- Consider a randomized study that includes a control arm and J treatment arms
- At a predetermined interim analysis (*the DTL look*), one of the treatment arms is selected based on the observed data and is continued until the end of the study
- Further analyses are conducted on *the selected arm* at information fractions $t_1 < t_2 < \ldots < t_K = 1$ to compare it with the control arm with respect to the primary endpoint

Drop-The-Losers Design ii



Notations

- W_j: A statistic based on the observed data at the DTL look that corresponds to arm j
- $\boldsymbol{W} = (W_1, \ldots, W_J)$
- $g(\boldsymbol{W})$: A arm-selection function with $g(\boldsymbol{W}) \in \{1, \ldots, J\}$
- $H_0^{(j)}$: The primary hypothesis to be tested at the final analysis

$$H_0^{(j)} : \text{The survival functions of arm } j \text{ and the control are identical}$$

- $Z_{j1}, \ldots, Z_{jK}$: Test statistics for $H_0^{(j)}$ when comparing treatment arm j versus control arm at information fractions $t_1, \ldots, t_K$, respectively
- $c_{j1}, \ldots, c_{jK}$: Group sequential design boundaries

Type I Error Inflation

- Assume that the group sequential boundaries $c_{j1}, \dots, c_{jK}$ control the type I error rate at α_j^* level
- Then, $\Pr(\Pi_j) \leq \alpha_j^*$ under $H_0^{(j)}$, where

$$\Pi_j = \{Z_{j1} > c_{j1}\} \cup \{Z_{j1} \leq c_{j1}, Z_{j2} > c_{j2}\} \cup \dots \\ \cup \{Z_{j1} \leq c_{j1}, \dots, Z_{j,K-1} \leq c_{j,K-1}, Z_{jK} > c_{jK}\}$$

- The rejection rate for $H_0^{(j)}$ is

$$r_j = \Pr(g(\boldsymbol{W}) = j, \Pi_j)$$

- The overall rejection rate is $r = \sum_{j=1}^{J} r_j$
- When $W_j \perp\!\!\!\perp (Z_{j1}, \dots, Z_{jK}) \ \forall\ j$, $r \leq \alpha^*$. What if dependant?

Example I i

- Consider $J = 2$ and $K = 1$
- Suppose PFS is the endpoint both at the DTL look and the final analysis
- Let Z_{j1} be the log-rank test statistic for $H_0^{(j)}$ and Z_{j0} be the test statistic at the DTL look
- Let $W_1 = Z_{10}$ and $W_2 = Z_{20}$
- Specify the arm-selection function

$$g(W_1 = Z_{10}, W_2 = Z_{20}) = \mathrm{I}(Z_{20} - Z_{10} - \Delta \leq 0) + 2\,\mathrm{I}(Z_{20} - Z_{10} - \Delta > 0)$$

- Under global null, the joint distribution of the log-rank test statistics $(Z_{10}, Z_{20}, Z_{11}, Z_{21})$ is a multivariate normal distribution with mean $\mathbf{0}$ and covariance

$$\Sigma = \begin{bmatrix} 1 & \frac{1}{2} & \sqrt{\frac{t_0}{t_1}} & \frac{1}{2}\sqrt{\frac{t_0}{t_1}} \\ \frac{1}{2} & 1 & \frac{1}{2}\sqrt{\frac{t_0}{t_1}} & \sqrt{\frac{t_0}{t_1}} \\ \sqrt{\frac{t_0}{t_1}} & \frac{1}{2}\sqrt{\frac{t_0}{t_1}} & 1 & \frac{1}{2} \\ \frac{1}{2}\sqrt{\frac{t_0}{t_1}} & \sqrt{\frac{t_0}{t_1}} & \frac{1}{2} & 1 \end{bmatrix}$$

- The family-wise error rate (FWER) is

$$r = r_1 + r_2 = \alpha^* + \int_{z_{\alpha^*}}^{\infty} \left[\Phi(d_1) - \Phi(d_2)\right] \phi(z) dz,$$

where

$$d_1 = \frac{2\Delta + z\sqrt{t_0/t_1}}{\sqrt{4 - t_0/t_1}},\ d_2 = \frac{2\Delta - z\sqrt{t_0/t_1}}{\sqrt{4 - t_0/t_1}}$$

- Note that $r > \alpha^*$ since $d_1 > d_2$, i.e., there is type I error inflation

- Suppose pCR is the endpoint at the DTL look with n patients per arm
- Let W_j be the observed pCR rate in arm j at the DTL look
- Under global null, let $q = E(W_1) = E(W_2)$
- Consider the same arm-selection function as in Example 1:

$$g(W_1, W_2) = \mathrm{I}(W_2 - W_1 - \Delta \leq 0) + 2\,\mathrm{I}(W_2 - W_1 - \Delta > 0)$$

- Let ρ be the correlation coefficient of W_j and Z_{j1}
- Then,

$$(W_1, W_2, Z_{11}, Z_{21}) \sim \text{MVN}\left(\begin{bmatrix} q \\ q \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \frac{q(1-q)}{n} & 0 & \rho\sqrt{\frac{q(1-q)}{n}} & 0 \\ 0 & \frac{q(1-q)}{n} & 0 & \rho\sqrt{\frac{q(1-q)}{n}} \\ \rho\sqrt{\frac{q(1-q)}{n}} & 0 & 1 & \frac{1}{2} \\ 0 & \rho\sqrt{\frac{q(1-q)}{n}} & \frac{1}{2} & 1 \end{bmatrix}\right)$$

Example II iii

- Subsequently, we can derive the FWER

$$r = \alpha^* + \int_{z_{\alpha^*}}^{\infty} \left[\Phi(d_1) - \Phi(d_2)\right] \phi(z) dz,$$

with

$$d_1 = \frac{\Delta/\sqrt{q(1-q)/n} + \rho z}{\sqrt{2-\rho^2}}, \ d_2 = \frac{\Delta/\sqrt{q(1-q)/n} - \rho z}{\sqrt{2-\rho^2}}$$

- When $\rho = 0$, $r = \alpha^*$
- When $\Delta = 0$, the FWER increases as ρ increases and is independent of n and q
- r monotonically increases as α^* increases

Correlation Coefficient ρ i

Proposition: *Denote* $\rho_{jk} = \mathrm{Cor}(W_j, Z_{jk})$. *Without censoring,*

$$\rho_{jk} \approx \sqrt{\frac{q\tau_k}{2(1-q)}} \left[\int_0^\infty \frac{S_1(t) f(t)}{S(t)} dt - 1 \right].$$

Under the proportional hazard assumption $S_1(t) = [S_0(t)]^\gamma$*,*

$$\rho_{jk} \approx \sqrt{\frac{q\tau_k}{2(1-q)}} \left(\frac{1}{\gamma} - 1 \right) E \left[\frac{1}{1 + \frac{q}{1-q} U^{1-\frac{1}{\gamma}}} \right],$$

where $U \sim \mathrm{Unif}(0,1)$.

(Partial) Proof: *Under the global null assumption, let* $T_{ij}|X_{ij} = x \sim f_x(t)$ *for* $x = 0, 1$*. Let* $S_x(t)$ *be the survival function for* $f_x(t)$*. Then,* T_{ij} *has marginal density function* $f(t) = f_1(t)q + f_0(t)(1-q)$*, and marginal survival function* $S(t) = S_1(t)q + S_0(t)(1-q)$*.*

$$
\begin{aligned}
\mathrm{Cor}(W_j, Z_{jk}) = \rho_{jk} &= \frac{\mathrm{Cov}\left(\sum_{i=1}^{n} X_{ij}/n, Z_{jk}\right)}{\sqrt{q(1-q)/n}} = \frac{\sum_{i=1}^{n} E(X_{ij}Z_{jk})}{\sqrt{nq(1-q)}} \\
&= \frac{\sum_{i=1}^{n} E(E(X_{ij}Z_{jk}|X_{ij}))}{\sqrt{nq(1-q)}} = \frac{\sum_{i=1}^{n} E(Z_{jk}|X_{ij}=1)\Pr(X_{ij}=1)}{\sqrt{nq(1-q)}} \\
&= \sqrt{\frac{nq}{1-q}} E(Z_{jk}|X_{1j}=1).
\end{aligned}
$$

Proposition: *With non-informative censoring, where the censoring time follows cumulative distribution function,* $1 - S_C(t)$,

$$\rho_{jk} \approx \frac{\sqrt{\frac{q\tau_k}{2(1-q)}} \int_0^\infty \left[\frac{S_1(t)f(t)}{S(t)} - f_1(t)\right] S_C(t)dt}{\sqrt{\int_0^\infty f(t)S_C(t)dt}}.$$

Under the proportional hazard assumption $S_1(t) = [S_0(t)]^\gamma$,

$$\rho_{jk} \approx \frac{\sqrt{\frac{q\tau_k}{2(1-q)}} \left(\frac{1}{\gamma} - 1\right) \int_0^\infty \frac{1}{1+\frac{q}{1-q}[S_1(t)]^{1-\frac{1}{\gamma}}} f_1(t)S_C(t)dt}{\sqrt{\int_0^\infty f(t)S_C(t)dt}}.$$

Correlation Coefficient ρ iv

Theorem: Under the proportional hazard assumption $S_1(t) = [S_0(t)]^\gamma$ *with* $\gamma \leq 1$, $\forall S_C(t)$,

$$\rho_{jk} \leq \sqrt{\frac{q\tau_k}{2(1-q)}} \left(\frac{1}{\gamma} - 1\right) \sqrt{E\left[\frac{1}{\left(1 + \frac{q}{1-q}U^{1-\frac{1}{\gamma}}\right)^2 \left(q + \frac{1-q}{\gamma}U^{\frac{1}{\gamma}-1}\right)}\right]},$$

where $U \sim \mathrm{Unif}(0, 1)$.

Type I Error Control i

- In summary, the DTL design considers the following parameters for type I error control:
  - n: number of patients in each arm at the DTL look,
  - q: response rate
  - $g(\boldsymbol{W})$: arm selection function, or more specifically the arm selection criteria, Δ
  - $\{t_k\}$: information fractions for the interim analyses
  - $\{\rho_{jk}\}$: correlations between W_j and Z_{jk},

Type I Error Control ii

- Given these design parameters $\Omega = \{n, q, \Delta, \{t_k\}, \{\rho_{jk}\}\}$, the task of type I error control becomes identifying

$$\widetilde{\alpha} = \sup_{\alpha^*} \{\alpha^* : r(\Omega) \leq \alpha\} ,$$

  where α^* is the type I error rate based on which the group sequential design boundary is derived
  - Roughly, "the remaining α" after the DTL look

Type I Error Control iii

- In practice, we propose to conservatively identify

$$\widetilde{\alpha} = \sup_{\alpha^*} \left\{ \alpha^* : \left\{ \sup_{q \in Q_q, \gamma \in Q_\gamma, \{\rho_{jk} = \rho^*_{jk}(\tau_k, q, \gamma)\}} r(\Omega) \right\} \leq \alpha \right\},$$

where

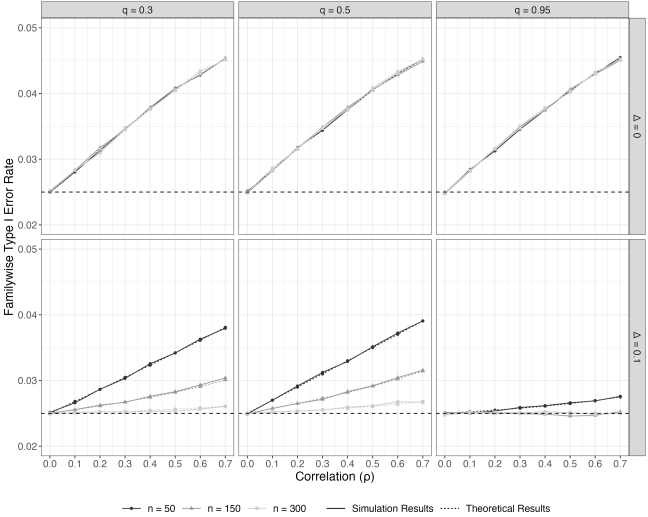
- Q_q and Q_γ are the parameter spaces of q and γ that are clinically meaningful, respectively
- $\rho^*_{jk}(\tau_k, q, \gamma)$ is the boundary for ρ_{jk}, given τ_k, q and γ

Numeric Studies

Type I Error Inflation i

- Correlation $\rho \in \{0, 0.1, \ldots, 0.7\}$
- pCR rate $q \in \{0.3, 0.5, 0.95\}$
- Sample size of each arm at the DTL look $n \in \{50, 150, 300\}$
- Consider $g(W_1, W_2) = \mathrm{I}(W_2 - W_1 - \Delta \leq 0) + 2\,\mathrm{I}(W_2 - W_1 - \Delta > 0)$
- $\Delta \in \{0, 0.1\}$
- Set $\alpha = 0.025$

Type I Error Inflation ii



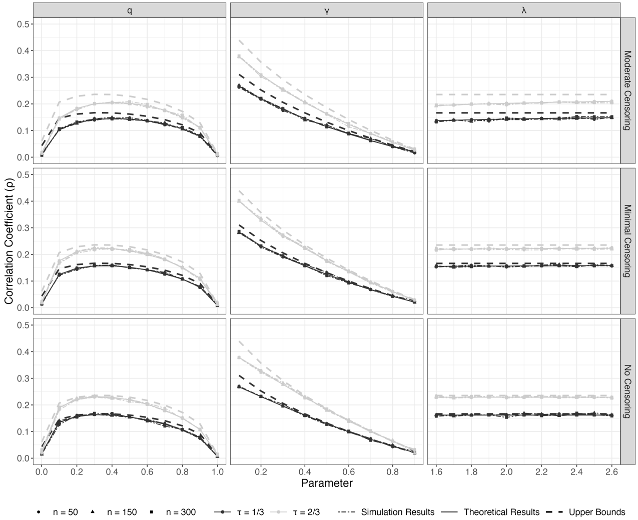
Type I Error Inflation iii

		$q = 0.30$		$q = 0.50$		$q = 0.95$	
Δ	ρ	$\widetilde{\alpha}$	FWER	$\widetilde{\alpha}$	FWER	$\widetilde{\alpha}$	FWER
0	0.1	0.0221	0.0250	0.0221	0.0250	0.0221	0.0253
	0.5	0.0151	0.0248	0.0151	0.0252	0.0151	0.0249
	0.7	0.0135	0.0249	0.0135	0.0248	0.0135	0.0248
0.1	0.1	0.0245	0.0251	0.0243	0.0252	0.0250	0.0249
	0.5	0.0221	0.0249	0.0213	0.0248	0.0250	0.0251
	0.7	0.0206	0.0252	0.0196	0.0251	0.0250	0.0250

Evaluating ρ i

- Sample size of each arm at the DTL look $n \in \{50, 150, 300\}$
- Total sample size of the selected arm $N = n/\tau$, with $\tau \in \{1/3, 2/3\}$
- For PFS, we assume $f_0(t) = \text{Exponential}(\lambda)$ for non-responders and $f_1(t) = \text{Exponential}(\gamma\lambda)$ for responders, with $\lambda \in [1.6, 2.6]$ and $\gamma \in [0.1, 0.9]$
- Censoring times following two different uniform distributions: $\text{Unif}(0, 2.2)$, referred to as *moderate censoring*, and $\text{Unif}(0, 6.5)$, referred to as *minimal censoring*
    - These distributions result in censoring rates ranging from 7% to 42%

Evaluating ρ ii



Application

A Hypothetical Trial

- Consider a hypothetical PD-1 and LAG-3 combination therapy study for patients with non-small cell lung cancer
- Consider two treatment arms, High Dose (HD) and Low Dose (LD), and compare them to the standard of care (SOC)

Initial Design

- Median PFS (mPFS) in the SOC arm: 180 days
- Hazard ratio between HD/LD and SOC: 0.6
- Annual drop-out rate: 5%
- Monthly enrollment rate: 20 patients per arm
- $N = 152$ per arm and a total of $D = 162$ events will achieve 90% power at a one-sided alpha level of 0.025

The DTL Look

- Based on *clinical, regulatory, and operational* considerations, the study plans to conduct a DTL analysis based on pCR when there are 80 patients in each arm (i.e., $n = 80$)
- If the pCR rate in the HD arm is more than 5% better than that in the LD arm (i.e., $\Delta = 0.05$), the study will drop the LD arm
- Otherwise, the study will drop the HD arm
- After the DTL look, an interim analysis for PFS will be conducted at an information fraction of 0.5
  - The O'Brien-Fleming boundaries will be used for the group sequential design

Type I Error Control

- First, quantify ρ by analyzing the data from the PD-1 arm of an existing study
  - The 95% confidence interval of the pCR rate was $[0.19, 0.32]$
  - The 95% confidence interval of hazard ratio between responders and non-responders was $[0.14, 0.34]$
- For simplicity, identify $\widetilde{\alpha}$ by ignoring the interim analysis
- Let $Q_q = [0.19, 0.32]$ and $Q_\gamma = [0.14, 0.34]$
- Obtain $\widetilde{\alpha} = 0.018$ through a grid search
- As a comparison, by naively assuming $\rho = 1$, the $\widetilde{\alpha}$ at the final analysis will be 0.013

Final Design and Operating Characteristics i

- Given the complexity of the study design, the sample size is fine-tuned by evaluating the study operating characteristics of the following scenarios with $\gamma = 0.15$

	pCR Rate (q)			mPFS in Days		
Scenario	SOC	LD	HD	SOC	LD	HD
I	0.2	0.2	0.5	180	180	300
II	0.2	0.4	0.5	180	276	300
III	0.2	0.5	0.5	180	300	300

- In each scenario, starting from $D = 162$ events, we conducted a grid search to ensure that the total number of events would provide 90% power

Final Design and Operating Characteristics

				Under H_1				Under H_0	
Sce.	P(HD)	Design	D	Rej. Any	Rej. LD	Rej. HD	Dur.	FWER	Dur.
I	1.000	Proposed	178	0.901	0.000	0.901	373	0.023	359
		$\rho = 1$	192	0.901	0.000	0.901	414	0.020	391
		Combination	197	0.900	0.000	0.900	418	0.016	404
II	0.723	Proposed	177	0.901	0.237	0.664	369	0.022	357
		$\rho = 1$	192	0.900	0.236	0.664	413	0.017	391
		Combination	193	0.900	0.237	0.664	406	0.018	394
III	0.259	Proposed	162	0.902	0.663	0.239	341	0.020	326
		$\rho = 1$	173	0.903	0.666	0.237	368	0.015	348
		Combination	173	0.905	0.669	0.237	359	0.015	348

- Considering all the information, the finalized design requires a total of 178 events for the final PFS analysis.

Software

R Package `dtlcor`

- An R package, `dtlcor`, has been developed and published on CRAN to implement the proposed DTL design
- `dtlcor` includes functions to calculate the correlation boundary ρ^* and $\widetilde{\alpha}$, and functions to conduct simulation studies to obtain design operating characteristics
- The package features *an R Shiny application* for interactive implementation of the design

Example

```
## install dtlcor package
install.packages("dtlcor")
require(dtlcor)

## get tilde alpha
alpha_t <- dtl_app_get_alpha_t(
    n = 80, N = 152,
    q_seq = seq(0.19, 0.32, 0.01),
    gamma_seq = seq(0.14, 0.34, 0.01),
    alpha = 0.025, delta = 0.05)

## simulate a single trial
trial <- dtl_app_sim_single(
    D = 162, n = 80, N = 152,
    mPFS = c(180, 276, 300),
    q = c(0.2, 0.4, 0.5),
    gamma = 0.15,
    drop_rate = 0.05, enroll = 20 * 12,
    interim_t = c(0.5, 1), delta = 0.05)
```

Discussion

Discussion i

- In this paper, we focus on using *intermediate or surrogate binary endpoints* in the DTL design for oncology trials
  - pCR is used as an illustrative example
  - Also applicable to binary outcomes, such as minimal residual disease status, circulating tumor DNA clearance, or tumor shrinkage status
- We derive theoretical formulas for the correlation coefficient based on practically reasonable assumptions
- More importantly, we derive a theoretical boundary
  - In all the simulation scenarios we have considered, the correlation has never exceeded 0.7

Discussion ii

- The proposal differs from the settings considered by alternative approaches
- Regarding the typical multi-arm multi-stage designs:
    - No hypothesis testing with respect to the surrogate endpoint
    - No comparison between treatment arms and control
    - Focus is the most conservative estimate (i.e., the upper bound) of the correlation
- Regarding the combination test
    - No clearly defined subset of control patients (or information) that corresponds to the p-value of the "first stage"

Discussion iii

- In summary, the proposed design will result in a smaller sample size and shorter study duration compared to the naive and most conservative approach
- Applying this correlation upper bound for multiplicity control will allow oncology studies using DTL designs to maintain well-controlled family-wise error rates, without being overly conservative

References

- Abbas, Rachid et al. (2022). **“A two-stage drop-the-losers design for time-to-event outcome using a historical control arm”.** In: *Pharmaceutical Statistics* 21.1, pp. 268–288.
- Bretz, Frank, José C Pinheiro, and Michael Branson (2005). **“Combining multiple comparisons and modeling techniques in dose-response studies”.** In: *Biometrics* 61.3, pp. 738–748.
- Budde, Michael and Peter Bauer (1989). **“Multiple test procedures in clinical dose finding studies”.** In: *Journal of the American Statistical Association* 84.407, pp. 792–796.
- Franchetti, Yoko, Stewart J Anderson, and Allan R Sampson (2013). **“An adaptive two-stage dose-response design method for establishing proof of concept”.** In: *Journal of biopharmaceutical statistics* 23.5, pp. 1124–1154.
- Posch, Martin et al. (2005). **“Testing and estimation in flexible group sequential designs with adaptive treatment selection”.** In: *Statistics in medicine* 24.24, pp. 3697–3714.
- Pozzi, L et al. (2013). **“A Bayesian adaptive dose selection procedure with an overdispersed count endpoint”.** In: *Statistics in Medicine* 32.28, pp. 5008–5027.
- Sampson, Allan R and Michael W Sill (2005). **“Drop-the-losers design: normal case”.** In: *Biometrical Journal: Journal of Mathematical Methods in Biosciences* 47.3, pp. 257–268.
- Schmidli, Heinz, Frank Bretz, and Amy Racine-Poon (2007). **“Bayesian predictive power for interim adaptation in seamless phase II/III trials where the endpoint is survival up to some specified timepoint”.** In: *Statistics in medicine* 26.27, pp. 4925–4938.
- Sill, Michael W and Allan R Sampson (2009). **“Drop-the-losers design: Binomial case”.** In: *Computational statistics & data analysis* 53.3, pp. 586–595.
- Wakana, Akira, Isao Yoshimura, and Chikuma Hamada (2007). **“A method for therapeutic dose selection in a phase II clinical trial using contrast statistics”.** In: *Statistics in medicine* 26.3, pp. 498–511.
- Wason, James et al. (2017). **“A multi-stage drop-the-losers design for multi-arm clinical trials”.** In: *Statistical methods in medical research* 26.1, pp. 508–524.
- Wason, James MS and Lorenzo Trippa (2014). **“A comparison of Bayesian adaptive randomization and multi-stage designs for multi-arm clinical trials”.** In: *Statistics in medicine* 33.13, pp. 2206–2221.