Optimizing trial designs for targeted therapies - A decision theoretic approach comparing sponsor and public health perspectives

Thomas Ondra    Sebastian Jobjörnsson

Section for Medical Statistics, Medical University of Vienna
Chalmers University of Technology

Seminar der Wiener Biometrische Sektion der Internationalen
Biometrischen Gesellschaft Region Österreich – Schweiz
Vienna, 11 February 2016

†Thomas Ondra, Sebastian Jobjörnsson, Robert Beckman,
Carl-Fredrik Burman, Franz König, Nigel Stallard and Martin Posch

This project has received funding from the European Union’s 7th Framework Programme for
research, technological development and demonstration under the IDEAL Grant Agreement no 602552, and the InSPIRe Grant Agreement no 602144.
Framework

Full Population $F$

Subgroup $S$

Complement $S' = F \setminus S$

- Overall treatment effect

$$\delta_F = \lambda \delta_S + (1 - \lambda) \delta_{S'}$$

where $\lambda$ is the prevalence of subgroup $S$.

- We assume $\delta_{S'} \leq \delta_S$.

- Allows for investigating the hypotheses $H_F : \delta_F \leq 0$ and $H_S : \delta_S \leq 0$. 
Enrichment Design: Randomize only patients of subgroup $S$ (say Biomarker $+\)$. Patients of the complement $S'$ are excluded from the trial (Biomarker $-$).

Classical Design: Recruit from the full population $F$. No Biomarker is determined.

Stratification Design: Include Biomarker $+$ and Biomarker $-$ patients. Stratify randomization by biomarker status.

- With the enrichment design one can test $H_S$, i.e., for a treatment effect in the subpopulation.
- With the classical design one can test $H_F$.
- With the stratification design one can test $H_S$ and $H_F$. 
Parallel group comparison of the means of normal distributions.

**Enrichment Design:**
- Test $H_S$ with a z-test.

**Classical Design:**
- $H_F$ with a z-test.

**Stratification Design:**
- Test $H_S$ and $H_F$ with a closed test, based on the Spiessens-Debois test for testing the global null hypothesis $H_F \cap H_S$ (Song and Chi, 2007; Spiessens and Debois, 2010).
Testing strategy in the stratified design

The stratified design allows for investigating $H_S$ and $H_F$.

Closed Testing principle

If $\eta = (\eta_{H_S}, \eta_{H_F}, \eta_{H_S \cap H_F})$ are local level alpha tests for $\mathcal{H} = \{H_S, H_F, H_S \cap H_F\}$, then the closed test $\psi_S = \min\{\eta_{H_S}, \eta_{H_S \cap H_F}\}$ and $\psi_F = \min\{\eta_{H_F}, \eta_{H_S \cap H_F}\}$ controls the FWER in the strong sense.

Local level $\alpha$ tests:
Test statistic for $H_S$: based on a z-test ($\eta_{H_S}$).
Test statistic for $H_F$: stratified z-test ($\eta_{H_F}$).
The test $\eta_{H_S \cap H_F}$ for the global null hypothesis $H_F \cap H_S$ will be based on the Spiessens–Debois test.
Testing the global null hypothesis $H_F \cap H_S$.

For adjusted significance levels $\alpha_F, \alpha_S$ we reject $H_F \cap H_S$ if

$$p_F \leq \alpha_F \quad \text{or} \quad p_S \leq \alpha_S,$$

where $p_F, p_S$ are the p-values of the z-tests for $H_F$ and $H_S$.

Some remarks:

- For fixed $\alpha_F$ and $\alpha$, the level $\alpha_S$ is chosen such that

$$\mathbb{P}_{H_F \cap H_S} (p_F < \alpha_F \text{ or } p_S < \alpha_S) = \alpha.$$

- For fixed $\alpha_F$ the level $\alpha_S$ increases with the prevalence $\lambda$ because the correlation of the test statistics increases.

- Formulas well known from group sequential tests.
A significant effect in $F$ might be totally driven by the effect in the subgroup.

To account for that we ask in addition to a significant effect in the full population that the effect in the complement (and the subgroup) show a positive trend.

Let $(\psi_F, \psi_S)$ denote the closed test based on the Spiessens-Debois test.

We define the modified testing procedure test via

\[
\tilde{\psi}_S := \psi_S \\
\tilde{\psi}_F := \psi_F \cdot 1_{\{p_S \leq \tau_S\}} \cdot 1_{\{p_{S'} \leq \tau_{S'}\}}
\]
Traditionally power arguments can be the basis for determining the best trial design.

An alternative is to apply a utility based approach (Graf et al., 2015; Beckman et al., 2011).

We model the sponsors/public health gain and costs of a particular trial design.

Best trial design is determined by maximizing the sponsors/public health’s profit.

In particular we optimize the following aspects of a clinical trial:

- Which type of design (Enrichment Design/Classical Design/Stratified Design) to choose?
- Which sample size?
- Which significance levels $\alpha_F$ and $\alpha_S$ for $H_F$ and $H_S$ in the weighted multiple test for the stratified design are optimal?
- Which thresholds $\tau_S, \tau_S'$ are optimal (optimized in public health view only).
The utility function

\[ U_k(d) = \tilde{\psi}_{F,d} \cdot \varphi_{F,d}^k + (1 - \tilde{\psi}_{F,d})\tilde{\psi}_{S,d} \cdot \varphi_{S,d}^k - C(d). \]

- \( k \in \{ \text{Sponsor, Public Health} \} \).
- \( \tilde{\psi}_{F,d} \) modified Spiessens-Debois test for \( H_F \) (= 0 for enrichment trials).
- \( \varphi_{F,d}^k \) measure of revenue if drug is licensed in \( F \).
- \( \tilde{\psi}_{S,d} \) modified Spiessens-Debois test for \( H_S \) (= 0 for classical trials).
- \( \varphi_{S,d}^k \) measure of revenue if drug is licensed in \( S \) only.
- \( C(d) \) cost for the trial.
The revenue measures

We assume that the revenue measures \( \varphi_{F,d}^{\text{Sponsor}}, \varphi_{S,d}^{\text{Sponsor}} \) depend on the data via the observed effect sizes \( \hat{\delta}_{F,d} \) and \( \hat{\delta}_{S,d} \):

\[
\varphi_{F,d}^{\text{Sponsor}} = N \cdot r_F \cdot (\hat{\delta}_{F,d} - \mu_F)^+ \\
\varphi_{S,d}^{\text{Sponsor}} = \lambda \cdot N \cdot r_S \cdot (\hat{\delta}_{S,d} - \mu_S)^+
\]

where

- \( N \) denotes the number of future patients (market size).
- \( r_F, r_S \) are revenue parameters.
- \( \mu_F, \mu_S \) denote clinically relevant thresholds.

The revenue measures for the public health view are given by

\[
\varphi_{F,d}^{\text{Public health}} = \varphi_F = N \cdot r_F \cdot (\delta_F - \mu_F) \\
\varphi_{S,d}^{\text{Public health}} = \varphi_S = \lambda \cdot N \cdot r_S \cdot (\delta_S - \mu_S)
\]
The costs of a trial are the same for the sponsor and the public health view.

- **Classical Trial**
  \[ C(d) = c_{\text{setup}} + 2nc_{\text{per-patient}}. \]

- **Stratified Trial**
  \[ C(d) = c_{\text{setup}} + c_{\text{Biomarker development}} + 2n(c_{\text{per-patient}} + c_{\text{screening}}). \]

- **Enrichment Trial**
  \[ C(d) = c_{\text{setup}} + c_{\text{Biomarker development}} + 2n(c_{\text{per-patient}} + \frac{c_{\text{screening}}}{\lambda}). \]
The optimal design is defined via:

\[ d^* \in \arg \max_{d \in D} E_\pi [U(d)], \]

where

\[ E_\pi[U(d)] = \int E_\Delta[U(d)] \pi(\Delta), \]
Plan:

1. How to compute the expected utilities?
2. Description of the different cases studied.
3. Presentation and discussion of plots for the different cases.
4. Some conclusions from the case studies.

Recall, the design parameters are:

- All designs: sample size.
- Stratified design: significance levels for the multiplicity adjustment procedure.
- Stratified design for public health view: the additional threshold for an effect in $S'$ that is required for approval in $F (\rho_{S'} < \tau_{S'})$. 
Expected utility for enrichment design

For the enrichment design $\tilde{\psi}_{F,E_n} = 0$, so that

$$U_S(E_n) = N\lambda r_S\psi_{S,E_n} \left(\hat{\delta}_{S,E_n} - \mu_S\right)^+ - C(E_n).$$

The expected utility given effect sizes $\Delta$ is

$$E[U_S(E_n)|\Delta] = N\lambda r_S \left((1 - \Phi(\kappa))(\delta_S - \mu_S) + \sqrt{\frac{2\sigma^2}{n}}\phi(\kappa)\right) - C(E_n),$$

$$\kappa = \sqrt{\frac{n}{2\sigma^2}} \left[\max \left(z_\alpha\sqrt{\frac{2\sigma^2}{n}}, \mu_S\right) - \delta_S\right].$$

Similarly, for the public health view

$$E[U_{PH}(E_n)|\Delta] = N\lambda r_S(\delta_S - \mu_S) \left(1 - \Phi \left(z_\alpha - \delta_S/\sqrt{\frac{2\sigma^2}{n}}\right)\right) - C(E_n).$$
For the classical design $\tilde{S}_C = 0$, so that

$$U_S(C_n) = Nr_F \psi_F, C_n \left( \hat{\delta}_F - \mu_F \right)^+ - C(C_n),$$

The expected utility given effect sizes $\Delta$ is

$$E[U_S(C_n) | \Delta] = Nr_F \left( (1 - \Phi(\kappa))(\delta_F - \mu_F) + \sqrt{V(\hat{\delta}_F)\phi(\kappa)} \right) - C(C_n),$$

$$V(\hat{\delta}_F) = \left( 2\sigma^2 + \lambda(1 - \lambda)(\delta_S - \delta_{S'})^2 \right) / n,$$

$$\kappa = \sqrt{V(\hat{\delta}_F)^{-1/2}} \left[ \max \left( z_\alpha \sqrt{2\sigma^2 / n}, \mu_F \right) - \delta_F \right].$$

Similarly, for the public health view

$$E[U_{PH}(C_n) | \Delta] = Nr_F(\delta_F - \mu_F) \left( 1 - \Phi \left( z_\alpha - \delta_F \sqrt{V(\hat{\delta}_F)^{-1/2}} \right) \right) - C(C_n).$$
The expected utility given the effect sizes $\Delta$ is given by

$$
E[U_S(S_{n,\alpha_S})|\Delta] = Nr_F E \left[ \tilde{\psi}_F \left( \hat{\delta}_F - \mu_F \right)^+ | \Delta \right] \\
+ Nr_S E \left[ \left( 1 - \tilde{\psi}_F \right) \tilde{\psi}_S \left( \hat{\delta}_S - \mu_S \right)^+ | \Delta \right] - C(S_{n,\alpha_S}).
$$

It can be computed using numerical integration, as can the corresponding expression for the public health view.
Scenarios and cases

Each particular situation is defined as follows:

1. Fix all parameters except the form of the prior, the market size and the biomarker costs.
2. Choose a scenario defining the form of the prior.
3. Choose a case defining the market size and the biomarker costs.
Fixed parameters for case studies

- Minimum sample size required by regulator: \( n_{\text{min}} = 50 \).
- Sample variance for each observation: \( \sigma = 1 \).
- One-sided significance level when testing: \( \alpha = 0.025 \).
- Minimal clinically relevant thresholds for regulatory approval:
  \( \mu_S = \mu_F = 0.1 \).
- Thresholds in the multiple test for the stratified design:
  \( \tau_S = \tau_{S'} = 0.3 \).
- Fixed setup cost for the trial: \( c_{\text{setup}} = 1 \text{ MUSD} \).
- Marginal cost per patient included: \( c_{\text{per-patient}} = 50\,000 \text{ USD} \).
We have considered priors $\pi_{\delta S, i, \delta S', i}$ on a grid $(\delta S, i, \delta S', i)$, $i = 1, \ldots, K$, of effect sizes.

**Scenario A** A point prior with $\pi_{\delta S, 0} = 1$ for $\delta S = 0.3$.

**Scenario B** A prior with $K = 3$. $\delta S = 0.3$ and $\pi_{\delta S, K-1} \frac{j}{\delta S} = \frac{1}{3}$, $j = 0, \ldots K - 1$.

**Scenario C** A point prior with $\pi_{\delta S, \delta S} = 1$ for $\delta S = 0.3$. 
Parameters for cases 1, 2 and 3

Reward and cost parameters in the utility function:

**Case 1** Large market and negligible biomarker costs.
\[ Nr_F = Nr_S = 10000 \text{ MUSD per unit of efficacy and} \]
\[ c_{\text{screening}} = c_{\text{Biomarker development}} = 0. \]

**Case 2** Small market and negligible biomarker costs.
\[ Nr_F = Nr_S = 1000 \text{ MUSD per unit of efficacy and} \]
\[ c_{\text{screening}} = c_{\text{Biomarker development}} = 0. \]

**Case 3** Small market with biomarker and screening costs.
\[ Nr_F = Nr_S = 1000 \text{ MUSD per unit of efficacy.} \]
\[ c_{\text{screening}} = 5000 \text{ USD per patient and} \]
\[ c_{\text{Biomarker development}} = 10 \text{ MUSD}. \]
Plots for cases 1, 2 and 3

For each case, we’ll look at

1. Optimal expected utility and sample size vs. $\lambda \in [0.04, 0.94]$.
2. Optimal significance levels of the multiple test for the stratified design, and the optimal threshold $\tau_{S'}$ vs. $\lambda \in [0.04, 0.94]$.
3. The power vs. $\lambda \in [0.04, 0.94]$.

Optimization of sample size is done for $n \leq 2000$. 
Optimal EU and Sample size (Case 1)

Sponsor

- Enrichment design
- Classical design
- Stratified design

Public Health

Sample Size

Utility (Mio. $)
Optimal sig levels for stratified design (Case 1)

Sponsor

- \( \lambda \)  
- Optimal significance levels

- Significance level for \( H_S \)
- Significance level for \( H_F \)

Public health

- \( \lambda \)  
- Optimal significance levels

- Significance level for \( H_S \)
- Significance level for \( H_F \)
- Optimized licensing Threshold
Power for the designs (Case 1)

Sponsor

Public health

Reject $H_S$

Reject $H_F$

Reject $H_S$ or $H_F$

Reject $H_S$ (Enrichment)

Reject $H_F$ (Classical)
Optimal EU and Sample size (Case 2)

**Sponsor**

- Enrichment design
- Classical design
- Stratified design

**Public Health**

Utility (Mio. $) vs. $\lambda$

Sample Size vs. $\lambda$
Optimal significance levels for stratified design (Case 2)

Sponsor

- Significant level for $H_S$
- Significant level for $H_F$

Public health

- Significant level for $H_S$
- Significant level for $H_F$
- Optimized licensing threshold
Power for the designs (Case 2)

Sponsor

Public health

Reject $H_S$

Reject $H_F$

Reject $H_S$ or $H_F$

Reject $H_S$ (Enrichment)

Reject $H_F$ (Classical)
Optimal EU and Sample size (Case 3)

Sponsor

- Utility (Mio. $)
- λ
- Enrichment design
- Classical design
- Stratified design

Public Health

- Utility (Mio. $)
- λ

Sample Size

- λ
Optimal significance levels for stratified design (Case 3)

**Sponsor**

- **Significance level for** $H_S$
- **Significance level for** $H_F$

**Public health**

- **Significance level for** $H_S$
- **Significance level for** $H_F$
- **Optimized licensing Threshold**
Power for the designs (Case 3)

Sponsor

Public health

Power of optimized designs

Reject $H_S$

Reject $H_F$

Reject $H_S$ or $H_F$

Reject $H_S$ (Enrichment)

Reject $H_F$ (Classical)
Optimality regions for the designs in the \((\delta_S, \lambda_S)-plane\)

Case 1

Case 2

Case 3

Sponsor

Public health
The optimal design depends heavily on the particular parameter configuration. However, our case studies indicates that

- Optimal expected utilities are larger for the sponsor.
- Optimal sample sizes are larger for the public health view.
- Since the public health decision maker optimises utility as a function of the true effects, it sometimes decides not to perform a study that a (commercial) sponsor would find attractive. Typically, this can be observed for priors corresponding to a belief in low effect sizes.
- Either the classical or stratified design tends to be optimal for the sponsor, while the enrichment design is sometimes optimal for the public health view.
- The relative size of the costs associated with the trial and the costs associated with biomarker testing has a strong impact on the optimality regions for the different designs.

