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

Thomas Ondra Sebastian Jobjörnsson

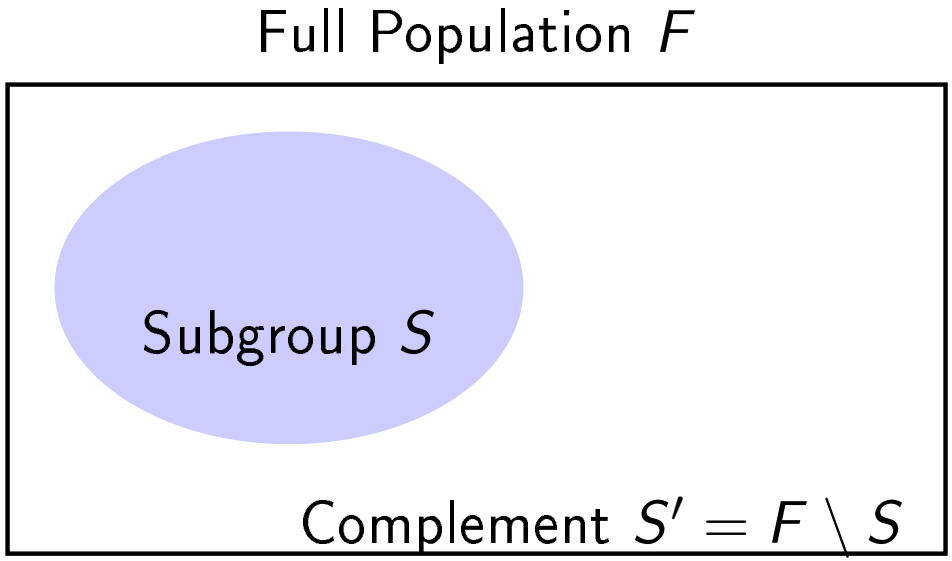
Section for Medical Statistics, Medical University of Vienna
Chalmers University of Technology

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



- Overall treatment effect

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

  where λ 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, classical and stratification design

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.

Testing problems

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 α tests:

Test statistic for H_S: based on a z-test (η_{H_S}).

Test statistic for H_F: stratified z-test (η_{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 α_F, α_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 α_F and α, the level α_S is chosen such that

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

- For fixed α_F the level α_S increases with the prevalence λ because the correlation of the test statistics increases.
- Formulas well known from group sequential tests.

Modified testing procedure

- 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 (ψ_F, ψ_S) denote the closed test based on the Spiessens-Debois test.
- We define the modified testing procedure test via

$$\tilde{\psi}_S := \psi_S$$
$$\tilde{\psi}_F := \psi_F \cdot \mathbf{1}_{\{p_S \leq \tau_S\}} \cdot \mathbf{1}_{\{p_{S'} \leq \tau_{S'}\}}$$

Optimizing trial designs

- 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 α_F and α_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

$$U_k(d) = \tilde{\psi}_{F,d} \cdot \varphi^k_{F,d} + (1 - \tilde{\psi}_{F,d})\tilde{\psi}_{S,d} \cdot \varphi^k_{S,d} - C(d).$$

- $k \in \{\text{Sponsor}, \text{Public Health}\}$.
- $\tilde{\psi}_{F,d}$ modified Spiessens-Debois test for H_F ($= 0$ for enrichment trials).
- $\varphi^k_{F,d}$ measure of revenue if drug is licensed in F.
- $\tilde{\psi}_{S,d}$ modified Spiessens-Debois test for H_S ($= 0$ for classical trials).
- $\varphi^k_{S,d}$ 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.
- μ_F, μ_S denote clinically relevant thresholds.

The revenue measures for the public health view are given by

$$\varphi_{F,d}^{\text{Public health}} = \varphi_F^{\text{Public health}} = N \cdot r_F \cdot (\delta_F - \mu_F)$$
$$\varphi_{S,d}^{\text{Public health}} = \varphi_S^{\text{Public health}} = \lambda \cdot N \cdot r_S \cdot (\delta_S - \mu_S)$$

The costs of a trial

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 trial design

The optimal design is defined via:

$$d^* \in \text{argmax}_{d \in D} E_\pi \left[U(d) \right],$$

where

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

Results - overview

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 ($p_{S'} < \tau_{S'}$).

Expected utility for enrichment design

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

$$U_{\mathrm{S}}(\mathtt{E}_n) = N\lambda r_S \psi_{S,\mathtt{E}_n} \left(\hat{\delta}_{S,\mathtt{E}_n} - \mu_S\right)^+ - C(\mathtt{E}_n).$$

The expected utility given effect sizes Δ is

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

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

Similarly, for the public health view

$$E[U_{\mathrm{PH}}(\mathtt{E}_n)|\Delta] = N\lambda r_S(\delta_S - \mu_S)\left(1 - \Phi\left(z_\alpha - \delta_S/\sqrt{\frac{2\sigma^2}{n}}\right)\right) - C(\mathtt{E}_n).$$

Expected utility for classical design

For the classical design $\tilde{\psi}_{S,\mathtt{c}_n} = 0$, so that

$$U_{\mathrm{S}}(\mathtt{c}_n) = N r_F \psi_{F,\mathtt{c}_n} \left(\hat{\delta}_F - \mu_F\right)^+ - C(\mathtt{c}_n),$$

The expected utility given effect sizes Δ is

$$E[U_{\mathrm{S}}(\mathtt{c}_n)|\Delta] = N r_F \left((1-\Phi(\kappa))(\delta_F - \mu_F) + \sqrt{V(\hat{\delta}_F)}\phi(\kappa)\right) - C(\mathtt{c}_n),$$

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

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

Similarly, for the public health view

$$E[U_{\mathrm{PH}}(\mathtt{c}_n)|\Delta] = N r_F(\delta_F - \mu_F)\left(1 - \Phi\left(z_\alpha - \delta_F V(\hat{\delta}_F)^{-1/2}\right)\right) - C(\mathtt{c}_n).$$

Expected utility for stratified design

The expected utility given the effect sizes Δ is given by

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

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: $c_{\text{setup}} = 1$ MUSD.
- Marginal cost per patient included: $c_{\text{per-patient}} = 50\,000$ USD.

Scenarios A, B and C

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 $\pi_{\delta_S, \frac{j}{K-1}\delta_S} = \frac{1}{3}$, $j = 0, \ldots K-1$.

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.
$Nr_F = Nr_S = 10\,000$ MUSD per unit of efficacy and
$c_{\text{screening}} = c_{\text{Biomarker development}} = 0$.

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

Case 3 Small market with biomarker and screening costs.
$Nr_F = Nr_S = 1000$ MUSD per unit of efficacy.
$c_{\text{screening}} = 5000$ USD per patient and
$c_{\text{Biomarker development}} = 10$ MUSD.

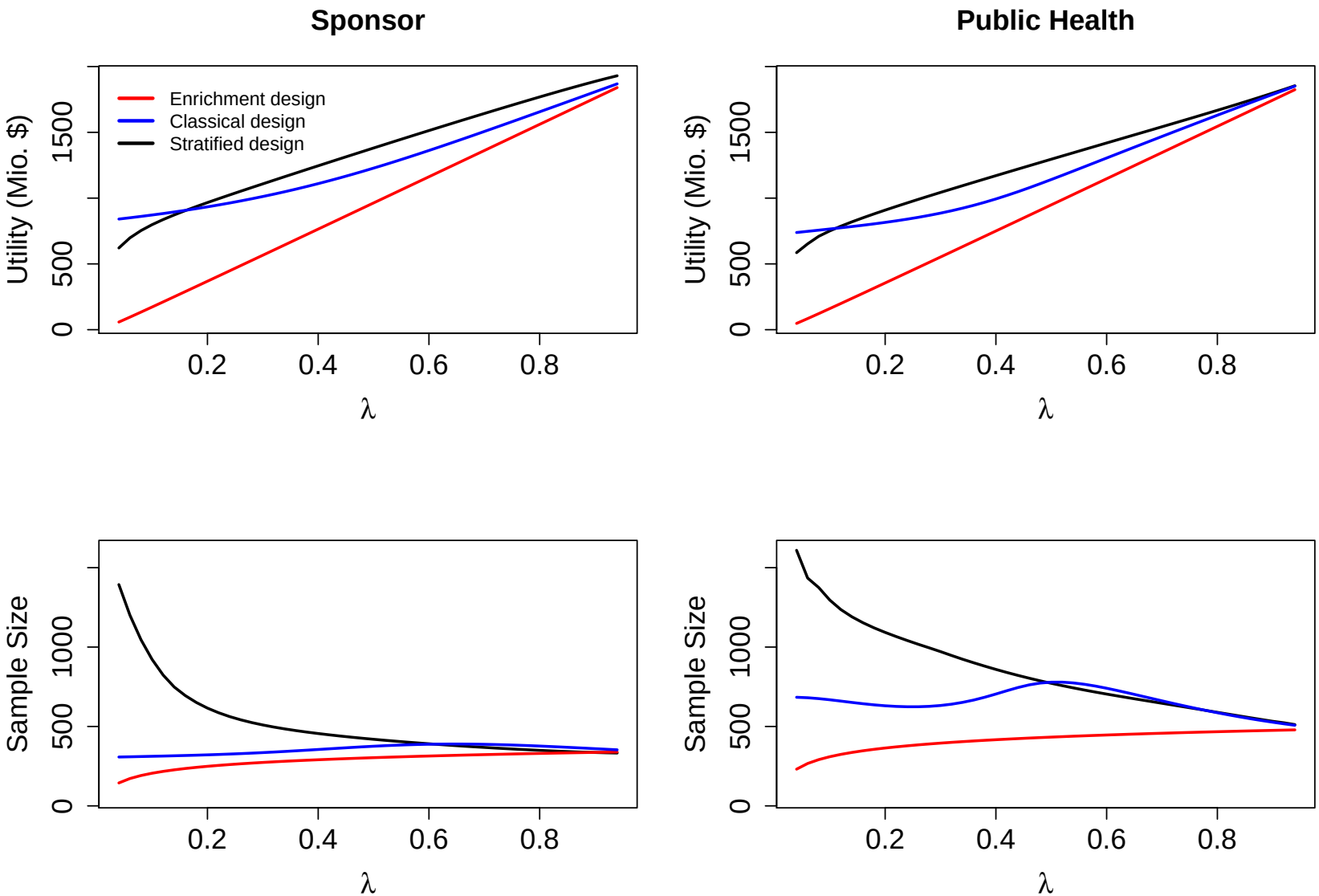
Plots for cases 1, 2 and 3

For each case, we'll look at

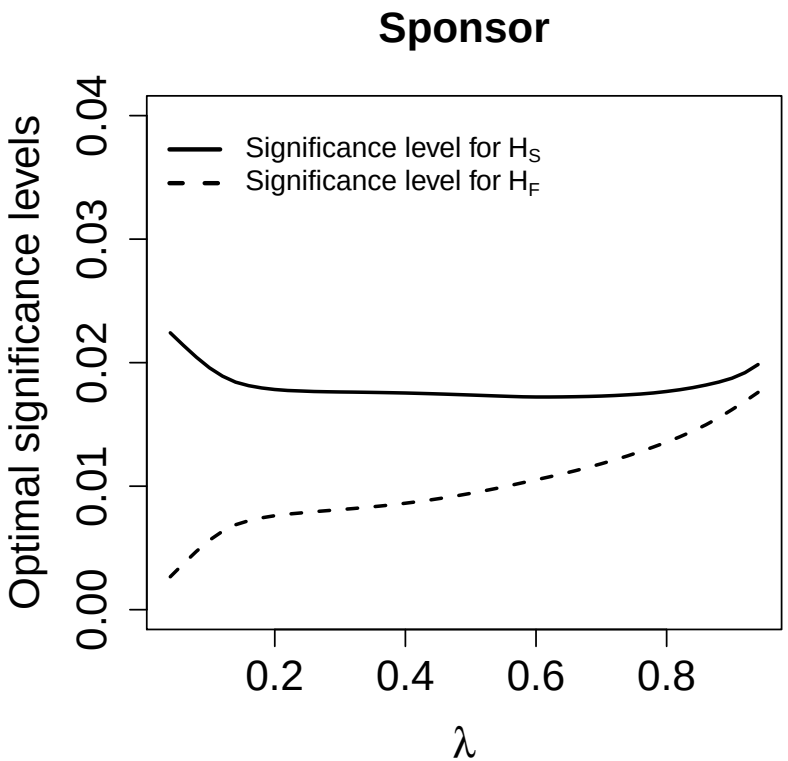
1. Optimal expected utility and sample size vs. $\lambda \in [0.04, 0.94]$.
2. Optimal significance levels of the multiple test for the stratified design, and the optimal threshold $\tau_{S'}$ vs. $\lambda \in [0.04, 0.94]$.
3. The power vs. $\lambda \in [0.04, 0.94]$.

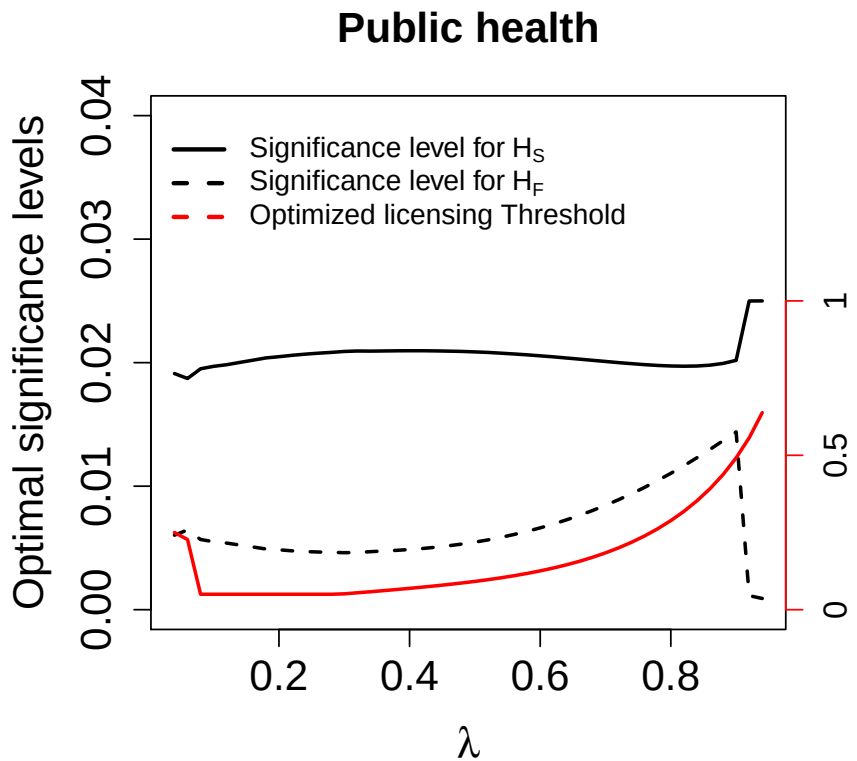
Optimization of sample size is done for $n \leq 2000$.

Optimal EU and Sample size (Case 1)

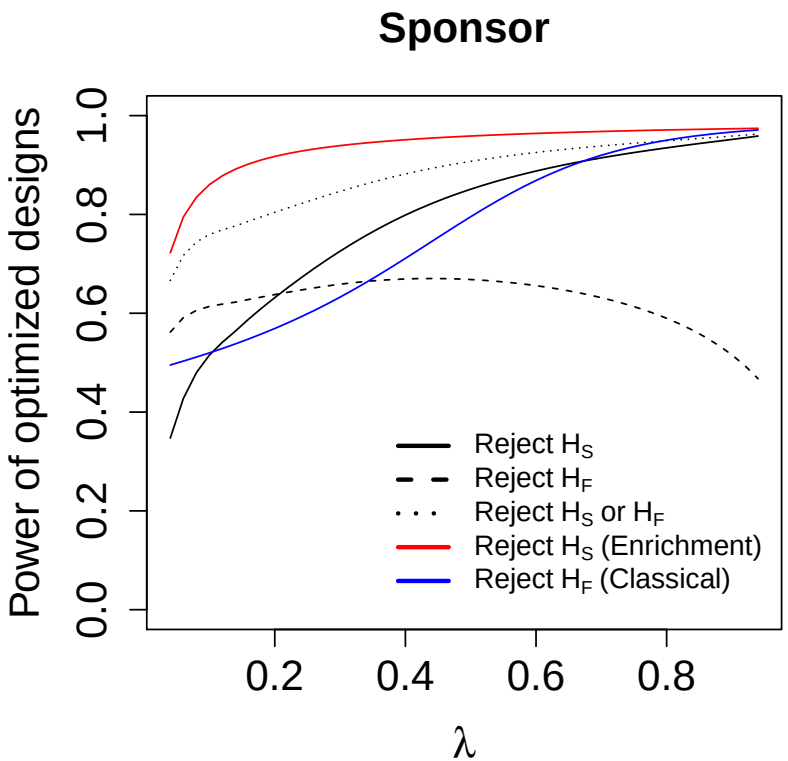


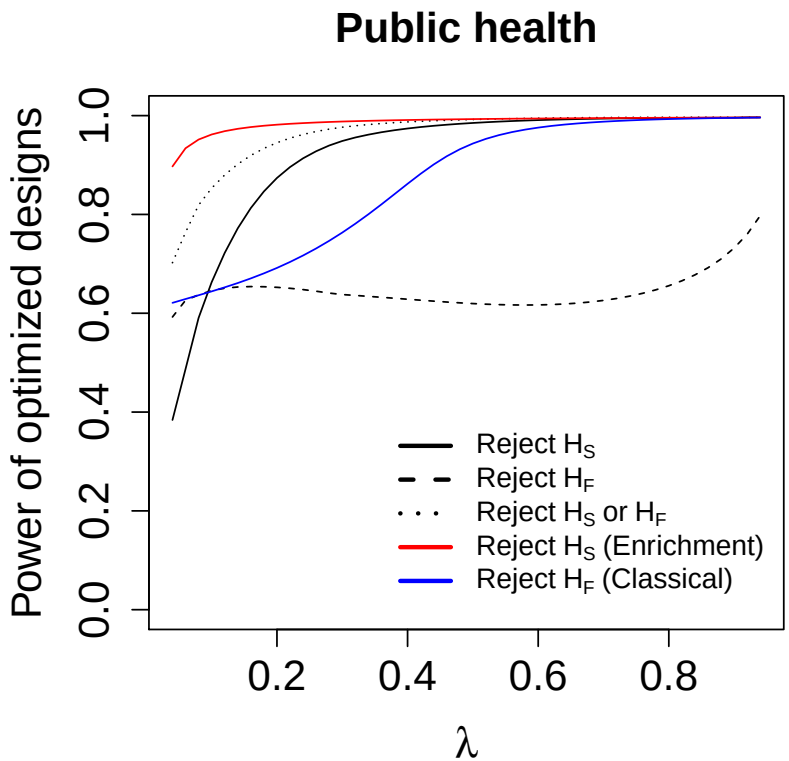
Optimal sig levels for stratified design (Case 1)



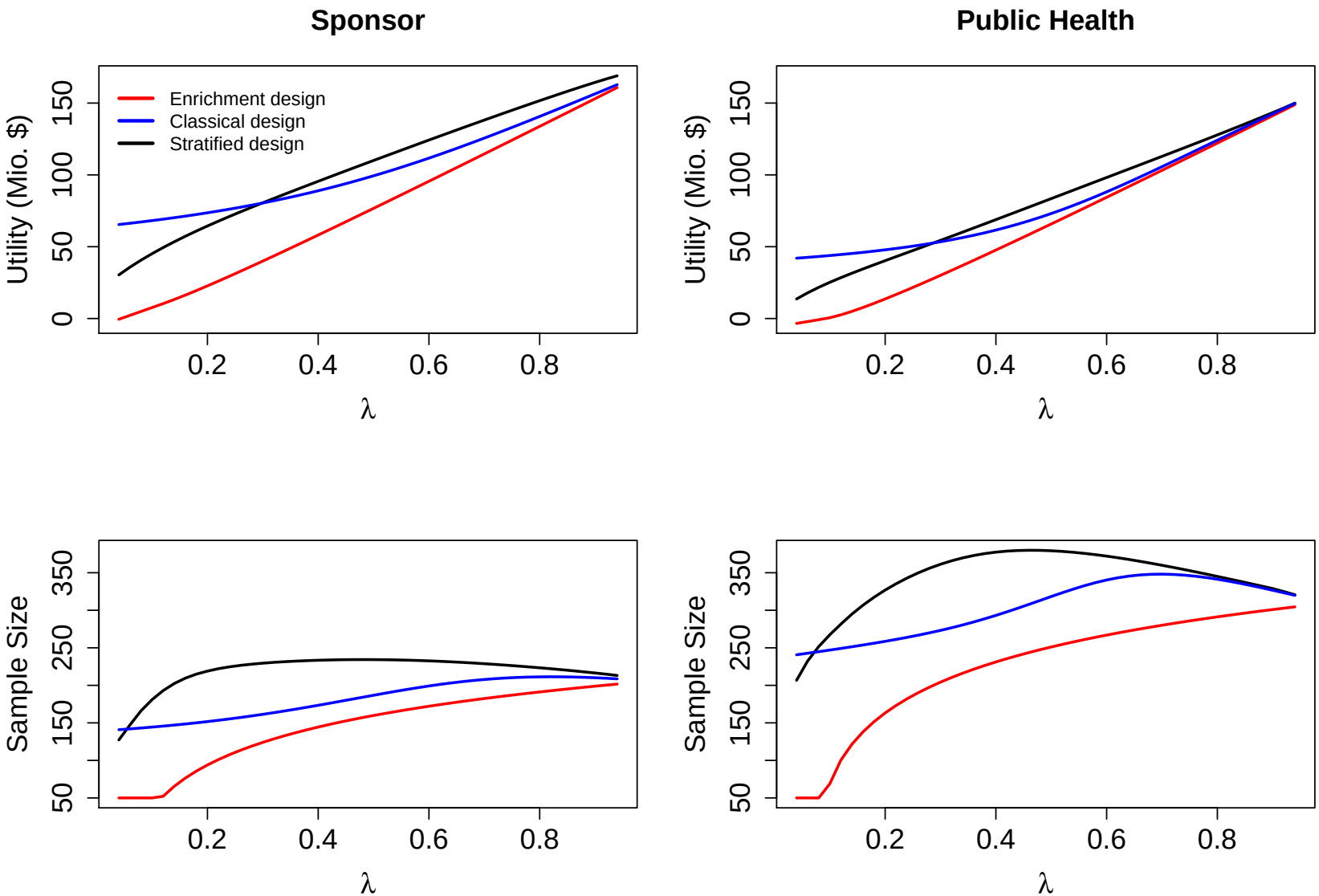


Power for the designs (Case 1)

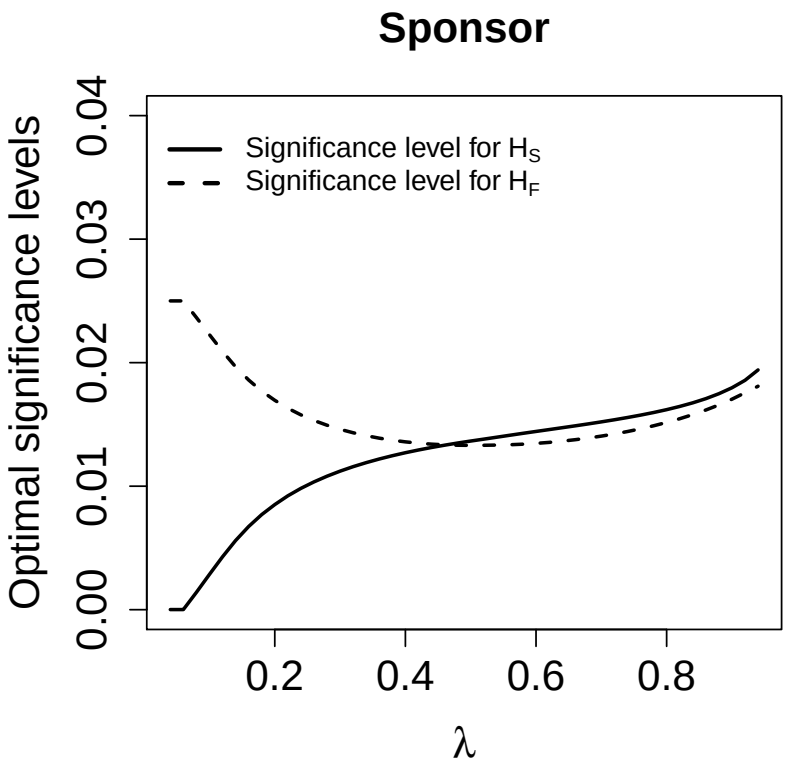


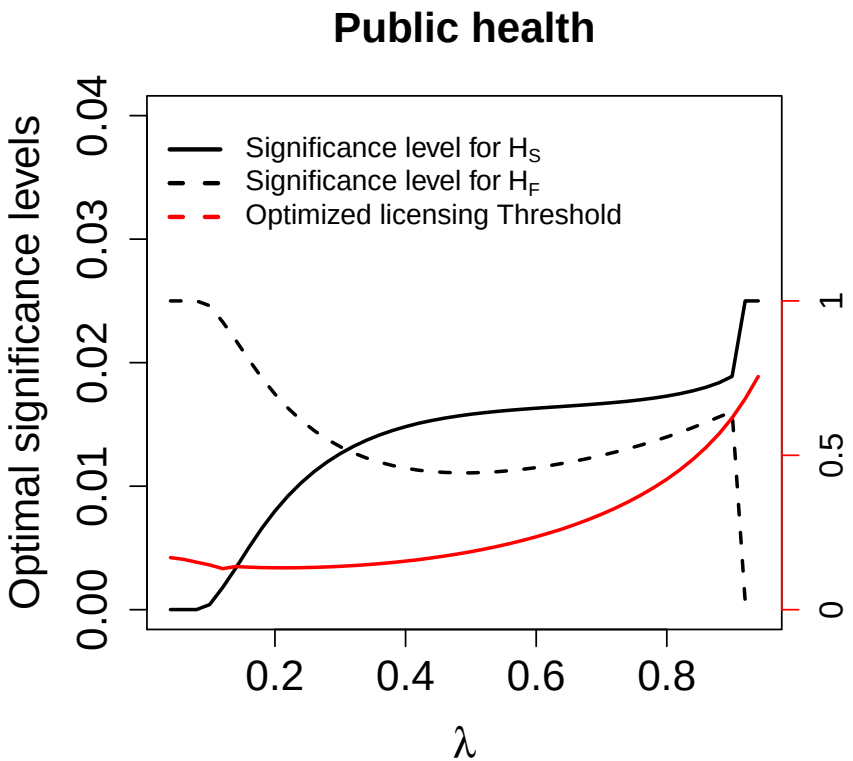


Optimal EU and Sample size (Case 2)

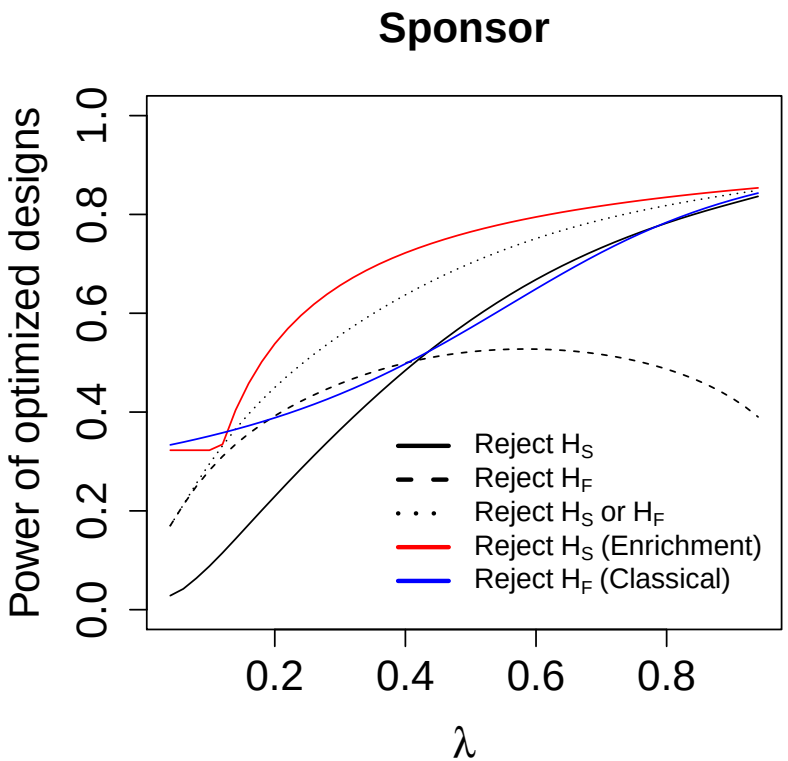


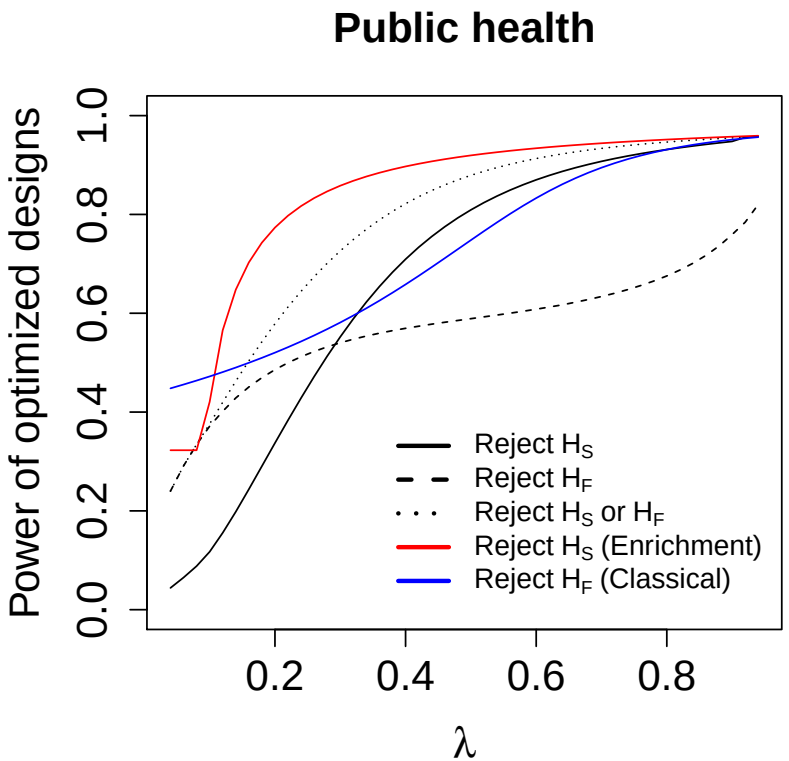
Optimal significance levels for stratified design (Case 2)



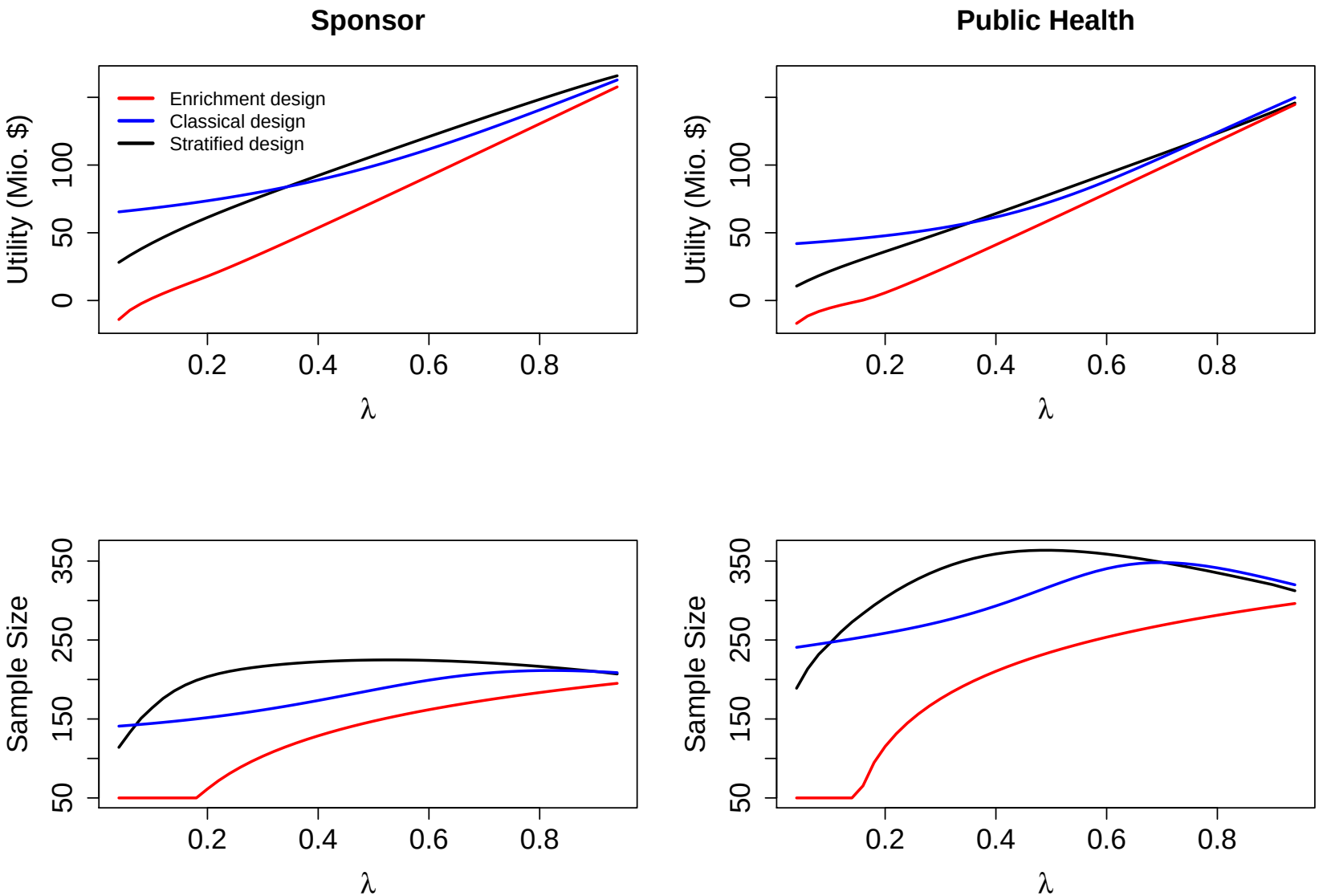


Power for the designs (Case 2)

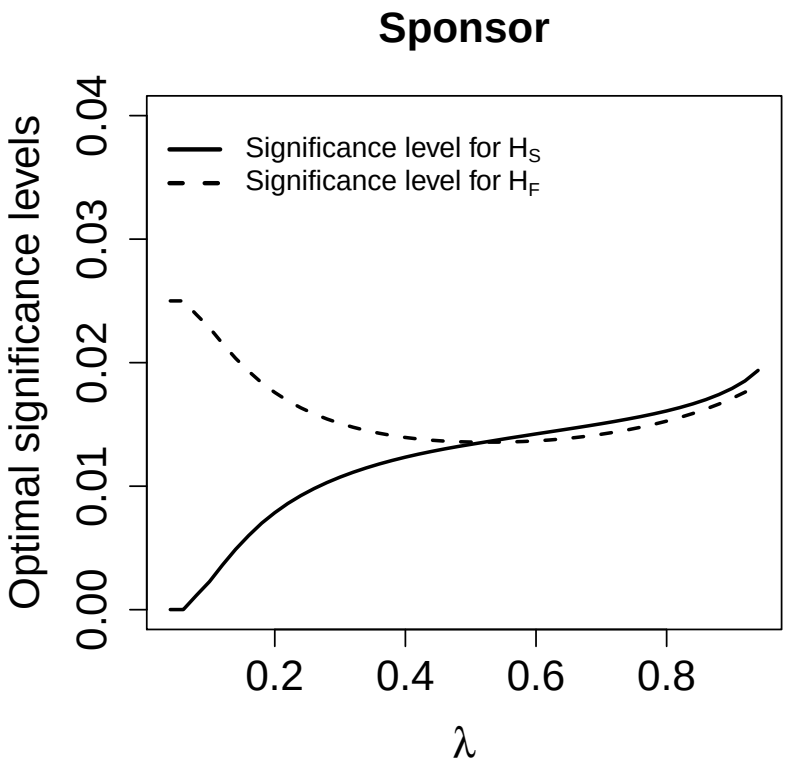


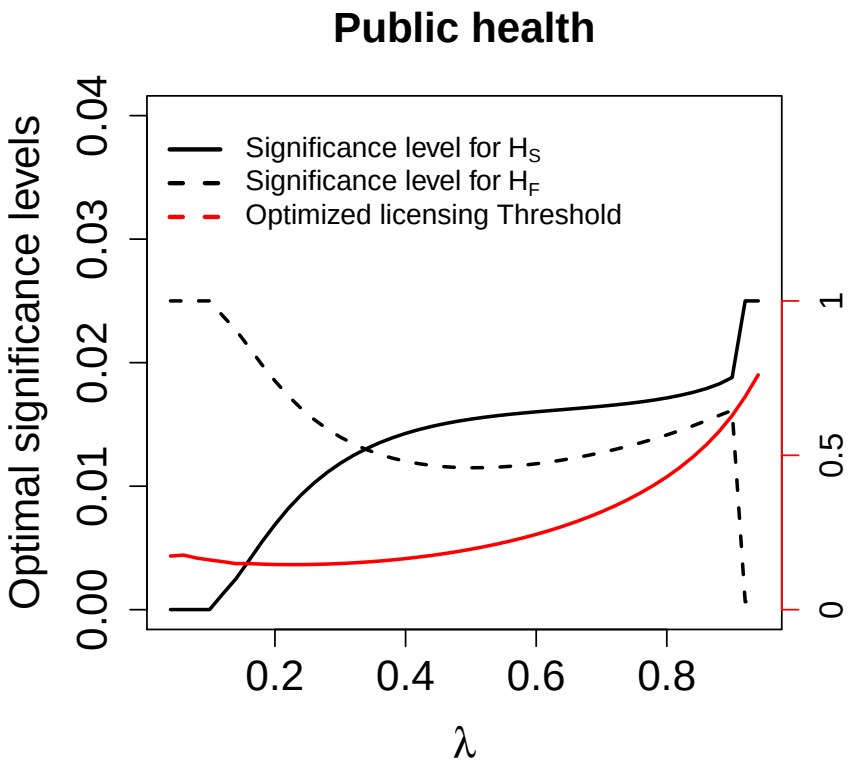


Optimal EU and Sample size (Case 3)

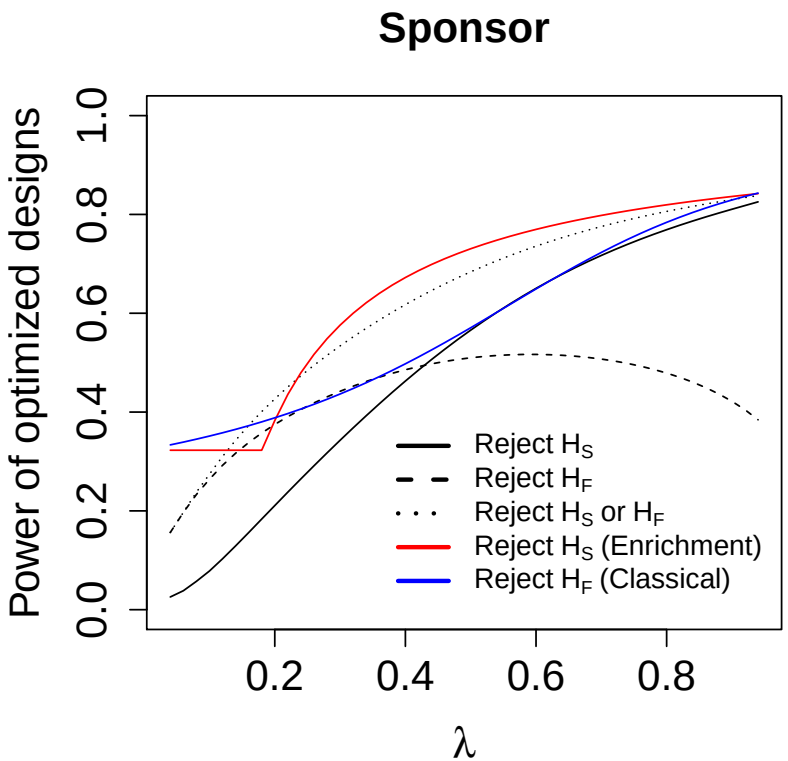


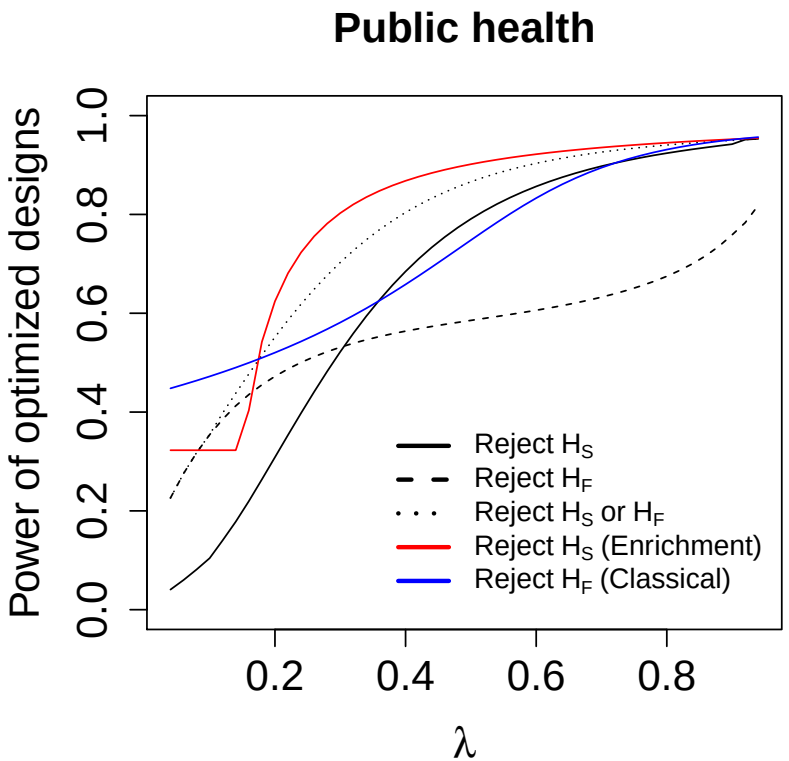
Optimal significance levels for stratified design (Case 3)



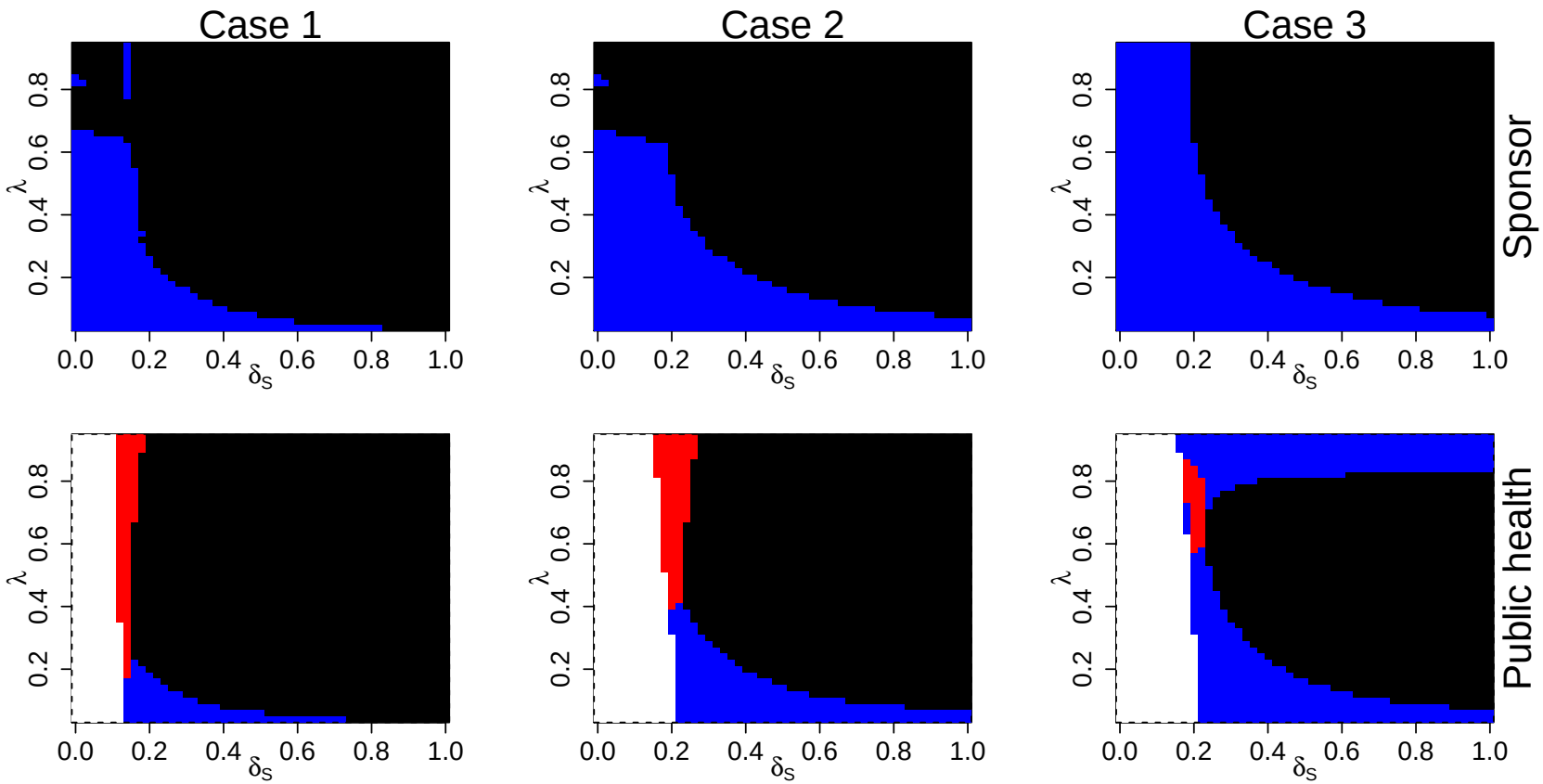


Power for the designs (Case 3)





Optimality regions for the designs in the (δ_S, λ_S)-plane



Summary of results from case studies

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.

References

Beckman, R. A., J. Clark, and C. Chen (2011). Integrating predictive biomarkers and classifiers into oncology clinical development programmes. *Nature Reviews Drug Discovery 10*(10), 735–748.

Graf, A. C., M. Posch, and F. Koenig (2015, January). Adaptive designs for subpopulation analysis optimizing utility functions. *Biometrical Journal 57*, 76–89.

Song, Y. and G. Y. H. Chi (2007, August). A method for testing a prespecified subgroup in clinical trials. *Statistics in Medicine 26*(19), 3535–3549.

Spiessens, B. and M. Debois (2010, November). Adjusted significance levels for subgroup analyses in clinical trials. *Contemporary Clinical Trials 31*(6), 647–656.