



Does Randomization protect against bias? What can be done to improve the level of clinical evidence of effectiveness

Ralf-Dieter Hilgers

Department of Medical Statistics, RWTH Aachen University
Coordinator of IDeAM Project

DKFZ, 2017, January 30th



FP7 HEALTH 2013 - 602552





- 1 The IDeAI Project
- 2 The Problem with Randomization
- 3 Evaluation of Randomization Procedures
- 4 Discussion
- 5 Outlook





New methodologies for clinical trials for small population groups FP7-HEALTH-2013-INNOVATION-1.

Objective develop new or improved statistical design methodologies for clinical trials aiming at the efficient assessment of the safety and/or efficacy of a treatment for small population groups in particular for rare diseases or personalised (stratified or individualised) medicine.

Multidisciplinary Framework involve all relevant stakeholders (including industry and patient advocacy groups) as appropriate. Ideally, results would lead to improvement of clinical trial guidelines. Collaboration with relevant organisations outside Europe is welcomed.

Expected Impact Cost efficient clinical trials deriving reliable results from trials in small population groups.



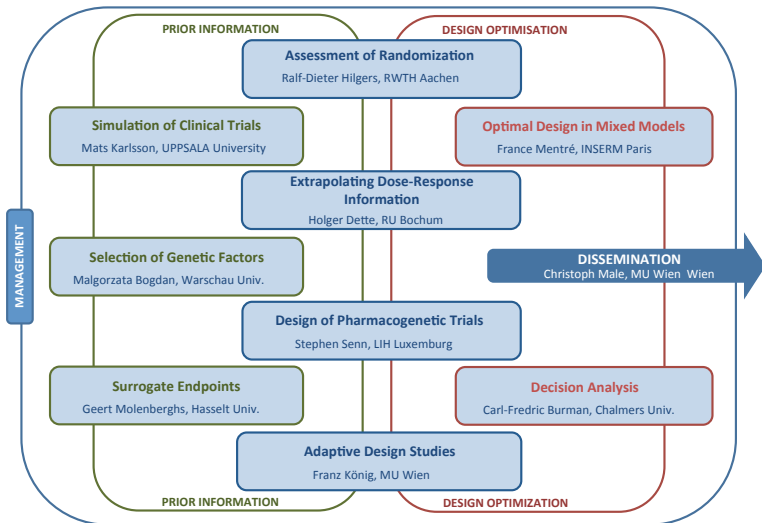


Integrated D^Esign and AnaL^Ysis of small population group trials

aims to refine the statistical methodology for clinical trials in small population groups by strictly following the concept of an improved integration of design, conduct and analysis of clinical trials from various perspectives.



Structure of the IDeAI Project



FP7 HEALTH 2013 - 602552





- ▶ **WP 2 (Randomization):** a new methodology for the selection of the best practice randomization procedure and subsequent analysis for a small population clinical trial taking possible bias into account
- ▶ **WP 3 (Extrapolation):** a new optimized design and analysis strategy for comparing dose response profiles to extrapolate clinical trial results from a large to a small population
- ▶ **WP 4 (Adaptive Design):** statistical methods to adapt the significance level and allow confirmatory decision-making in clinical trials with vulnerable, small populations
- ▶ **WP 5 (Optimal Design):** design evaluation methods enabling small clinical trials to be analyzed through modeling of continuous or discrete longitudinal outcomes.
- ▶ **WP 6 (Pharmacogenetic):** approaches to planning and analyzing trials for identifying individual response and examining treatment effects in small populations





- ▶ **WP 7 (Simulation):** new methods for sample size calculation, type 1 error control, model averaging and parameter precision in small populations group trials within non-linear mixed effects modelling
- ▶ **WP 8 (Genetic factors):** new methods for identifying biomarkers and prognostic scores based on high dimensional genetic data in small population group trials
- ▶ **WP 9 (Decision Analysis):** how to optimize the overall value of drug development to patients, to regulators and to society under opacity in regulatory and payer rules as well as in very rare diseases
- ▶ **WP 10 (Surrogate Endpoints):** methodology to evaluate potential surrogate markers and to analyze data from a small numbers of small trials, with emphasis on fast and easy computational strategies

`www.IDeAI.rwth-aachen.de`

`www.IDeAI.rwth-aachen.de/?page_id=806#toggle-id-12`





- What the theory tells us:
 - ▶ not any randomization procedure performs best with all criteria, Rosenberger (2016), Atkinson (2014)
- What applied scientist mostly feel:
 - ▶ scepticism to randomization
 - ▶ do not well understood randomization principle
 - ▶ is just allocation and think unequal group size is a major problem
 - ▶ think that randomization is for balancing covariates but does mostly not work
 - ▶ select a procedure by opinion or software availability
- What the literature mirrors:
 - ▶ no training in randomization
 - ▶ no recommendation to give scientific arguments for the choice of randomization procedure, neither ICH Guidelines nor CONSORT Statement

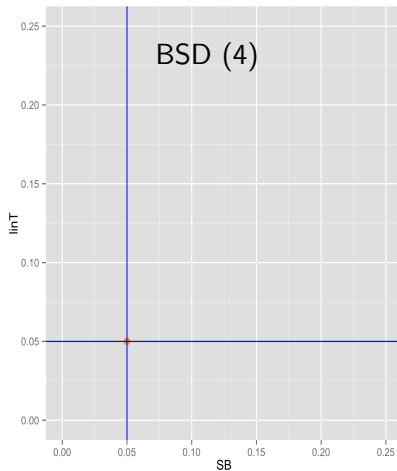
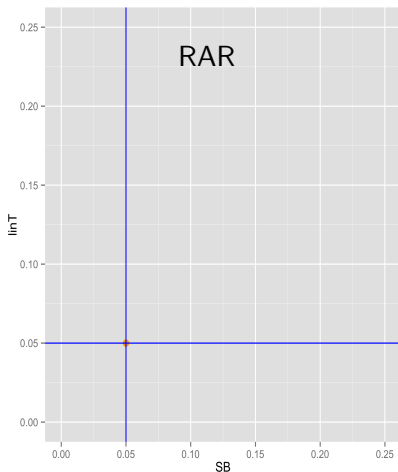


Impact of Bias on Type I Error Probability (N=96)



setting: $N_E = N_C = 48, \eta = 0.0 \times \text{effectsize } (\delta), \theta = 0.0 \times \sigma$

SB: Selection Bias; linT: Linear Time Trend



FP7 HEALTH 2013 - 602552

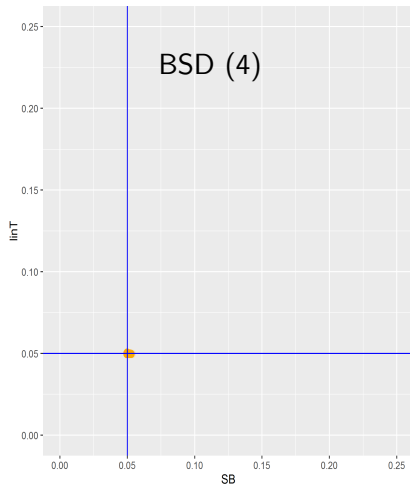
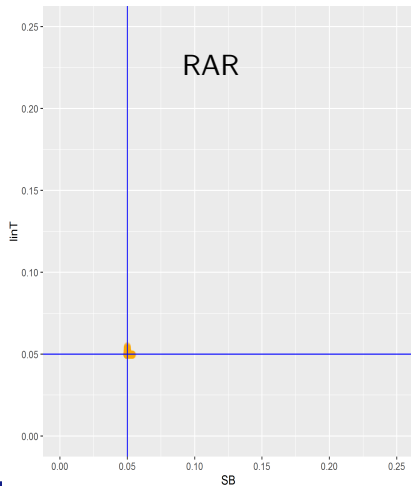


Impact of Bias on Type I Error Probability (N=96)



setting: $N_E = N_C = 48, \eta = 0.1 \times \delta, \theta = 0.2 \times \sigma$

SB: Selection Bias; linT: Linear Time Trend



FP7 HEALTH 2013 - 602552

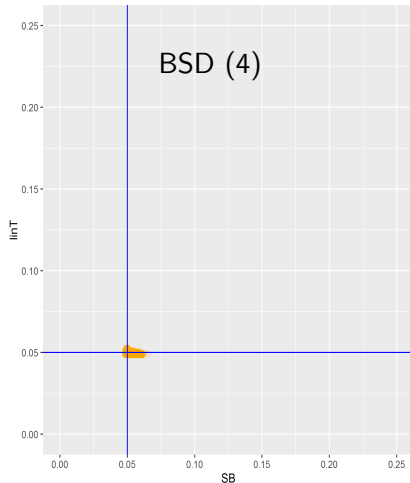
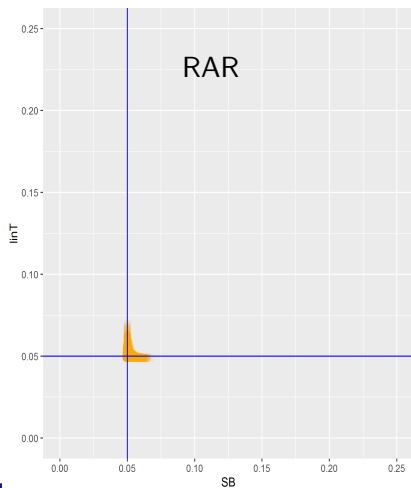


Impact of Bias on Type I Error Probability (N=96)



setting: $N_E = N_C = 48, \eta = 0.2 \times \delta, \theta = 0.4 \times \sigma$

SB: Selection Bias; linT: Linear Time Trend



FP7 HEALTH 2013 - 602552

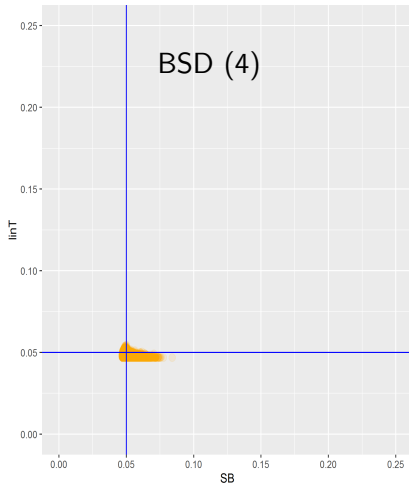
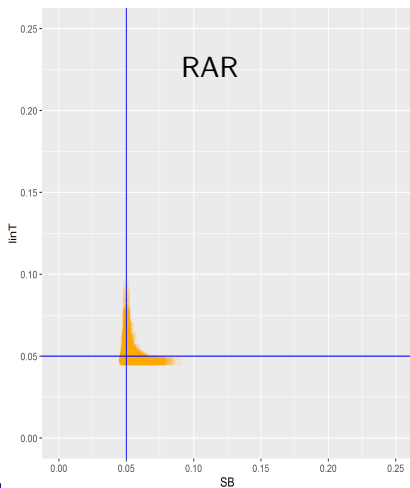


Impact of Bias on Type I Error Probability (N=96)



setting: $N_E = N_C = 48, \eta = 0.3 \times \delta, \theta = 0.6 \times \sigma$

SB: Selection Bias; linT: Linear Time Trend



FP7 HEALTH 2013 - 602552

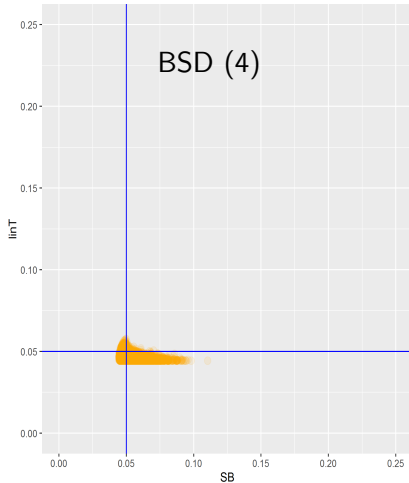
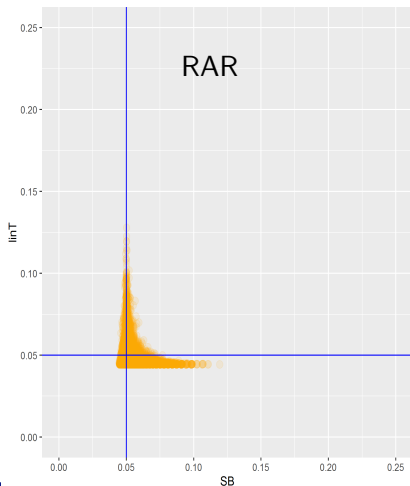


Impact of Bias on Type I Error Probability (N=96)



setting: $N_E = N_C = 48, \eta = 0.4 \times \delta, \theta = 0.8 \times \sigma$

SB: Selection Bias; linT: Linear Time Trend



FP7 HEALTH 2013 - 602552

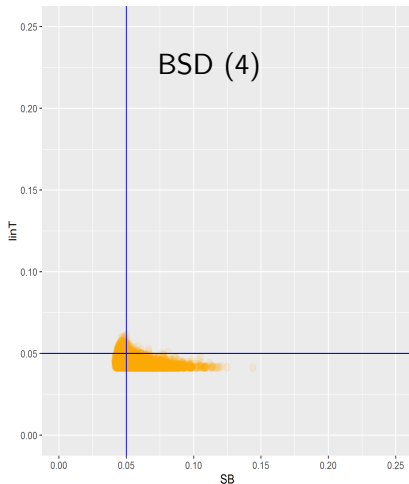
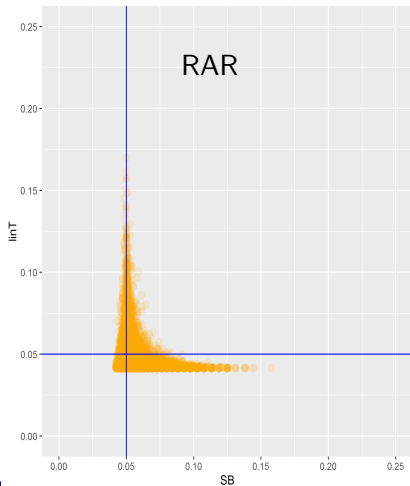


Impact of Bias on Type I Error Probability (N=96)



setting: $N_E = N_C = 48, \eta = 0.5 \times \delta, \theta = 1.0 \times \sigma$

SB: Selection Bias; linT: Linear Time Trend



FP7 HEALTH 2013 - 602552





Empirical type-I-error probability of a two sided t-test

N	$\delta(N)$	BSD (2)	CR	EBCD ($\frac{2}{3}$)	MP(2)	PBR(4)	RAR
8	2.381	0.064	0.058	0.089	0.118	0.141	0.102
20	1.325	0.075	0.054	0.093	0.129	0.177	0.082
32	1.024	0.083	0.055	0.097	0.137	0.188	0.072
40	0.909	0.088	0.053	0.100	0.140	0.195	0.071

- $N_E = N_C, N_E + N_C = N$
- $\delta(N) : \alpha = 0.05, 1 - \beta = 0.8$
- selection bias effect $\eta = \frac{\delta(N)}{2}$

using R with 100 000 replications





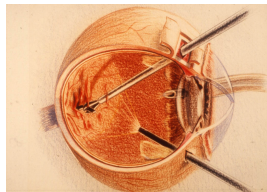
Evaluation of Randomization Procedures for Trial Design Optimization

- 1 **Introduction** - intend select the best practice randomization procedure (RP) to improve the level of evidence
- 2 **Objective** - select a best practice RP
- 3 **ERDO framework**
 - ▶ **Assumptions** - incl. design, clinical setting
 - ▶ **Options** - suitable set of RP's
 - ▶ **Metrics** - evaluation criterion e.g. averaged (empirical) type I error rate
- 4 **Evaluation Methods** - incl. statistical model, software, presentation of results, decision rule
- 5 **Result and Decision**
- 6 **Discussion and Clinical implication** - select the best practice (RP)
- 7 **Conclusion** choice of randomization design

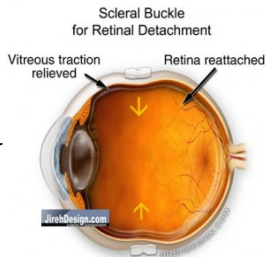




- scleral buckling (SB) with primary pars plana vitrectomy (PPV) in rhegmatogenous retinal detachment (SPR-Study, Heimann 2007)



- Additional encircling band might improve one year best corrected visual acuity results in the scleral buckling group.



<http://www.retinaeyedoctor.com/tag/eye/>

FP7 HEALTH 2013 - 602552





Primary Endpoint

- Change in Best Corrected Visual Acuity (BCVA) one year after surgery to baseline

Clinical Trial Layout

- parallel group design
- targeted allocation ratio 1:1, with a fixed sample design

Sample Size Calculation from SPR study data

- Sample size: 65 patients per group (VA: 0.52 (SD 0.77) with 0.90 (SD 0.73) without encircling band, t-test, two-sided 5% significance level, power of 80%, pooled standard deviation 0.765), with a loss in power of 5% if the allocation is at most 2:1 (88 to 44).





two arm parallel group design, continuous endpoint

Aim: test the hypotheses $H_0 : \mu_E = \mu_C$ vs. $H_1 : \mu_E \neq \mu_C$

Model for two arm parallel group design with continuous endpoint

$$Y_i = \mu_E T_i + \mu_C(1 - T_i) + \tau_i + \epsilon_i, \quad 1 \leq i \leq N_E + N_C$$

- allocation

$$T_i = \begin{cases} 1 & \text{if patient } i \text{ is allocated to group } E \\ 0 & \text{if patient } i \text{ is allocated to group } C \end{cases}$$

- μ_j expected response under treatment $j = C, E$
- τ_i denotes the fixed unobserved "bias" effect acting on the response of patient i
- errors ϵ_i iid $\mathcal{N}(0, \sigma^2)$





two arm parallel group trial continuous endpoint

Aim: test the hypotheses $H_0 : \mu_E = \mu_C$ vs. $H_1 : \mu_E \neq \mu_C$

use t-Test (under misspecification)

$$S = \frac{\sqrt{\frac{N_E N_C}{N_E + N_C}} (\tilde{y}_E - \tilde{y}_C)}{\frac{1}{N_E + N_C - 2} \left(\sum_{i=1}^N T_i (y_i - \tilde{y}_E)^2 + \sum_{i=1}^N (1 - T_i) (y_i - \tilde{y}_C)^2 \right)} \sim t_{N_E + N_C - 2, \vartheta, \lambda}$$

$$\text{where } \tilde{y}_E = \frac{1}{N_E} \sum_{i=1}^N y_i T_i ; \quad \tilde{y}_C = \frac{1}{N_C} \sum_{i=1}^N y_i (1 - T_i) ; \quad N = N_E + N_C$$





Theorem: Under $H_0 : \mu_E = \mu_C$ the type-I-error probability for the two arm parallel group normal model (under misspecification) for the allocation sequence $\mathbf{T} = (T_1, \dots, T_{N_E+N_C})$ is

$$P(|S| > t_{N_E+N_C-2}(1-\alpha/2) | \mathbf{T}) \\ = F_{N-2, \vartheta, \lambda}(t_{N_E+N_C-2}(\alpha/2)) + 1 - F_{N_E+N_C-2, \vartheta, \lambda}(t_{N_E+N_C-2}(1-\alpha/2)).$$

$F_{N_E+N_C-2, \vartheta, \lambda}$ denotes the distribution function of the doubly non-central t-distribution with $N_E + N_C - 2$ degrees of freedom and parameters

$$\vartheta = \frac{1}{\sigma} \sqrt{\frac{N_E N_C}{N_E + N_C}} (\tilde{\tau}_E - \tilde{\tau}_C) \quad \lambda = \frac{1}{\sigma^2} \left[\sum_{i=1}^N \tau_i^2 - N_E \tilde{\tau}_E^2 - N_C \tilde{\tau}_C^2 \right]$$

where $\tilde{\tau}_E = \frac{1}{N_E} \sum_{i=1}^N \tau_i T_i$; $\tilde{\tau}_C = \frac{1}{N_C} \sum_{i=1}^N \tau_i (1 - T_i)$





two arm parallel group trial continuous endpoint

Biasing policy according to convergence strategy

$$\tau_i = \begin{cases} \eta & \text{if } n_E(i-1) < n_C(i-1) \\ 0 & \text{if } n_E(i-1) = n_C(i-1) \\ -\eta & \text{if } n_E(i-1) > n_C(i-1) \end{cases}$$

- η proportional to effect size δ
- $\tau_i = \eta [\text{sign}(n_E(i-1) - n_C(i-1))]$
- $n_j(i)$: assignments to treatment j after i allocations

(Proschan 1994)

(Kennes 2011)





two arm parallel group trial continuous endpoint

Biasing policy according to convergence strategy

$$\tau_i = \theta \times \begin{cases} \frac{i}{N_E + N_C} & \text{linear time trend} \\ \mathbb{1}_{i \geq S(i)} & \text{stepwise trend} \\ \log\left(\frac{i}{N_E + N_C}\right) & \text{log trend} \end{cases}$$

- θ proportional to variance
- other functions are possible
- long recruitment time in rare diseases, (EMA, 2006)
 - ▶ changes in population characteristics
 - ▶ learning effect in therapy / surgical experience (Hopper, 2007)
 - ▶ change in diagnosis (FDA, 2011), etc.
- special form of accidental bias, when considering a time-heterogeneous covariate (Tamm, 2014)





two arm parallel group trial continuous endpoint

Joint Additive Bias

$$\tau_i = \underbrace{\theta \frac{i}{N_E + N_C}}_{\text{time trend}} + \underbrace{\eta [\text{sign}(n_E(i-1) - n_C(i-1))]}_{\text{selection bias}}$$

- weighted additive (selection and chronological) bias model
- weights via definition of θ and η
- multiplicative could also be done
- different shape of time trend can be incorporated (Tamm, 2014)
- relaxed version of bias policy (non strict decision, random η)





Bias effects were estimated from SPR-data and expressed as portion of the effect size.

Selection effect from SPR study data

- Selection bias of a reasonable 15% of the maximal treatment effect (i.e. $\eta = 0.08$)

from two way ANOVA with main effects

Time trend from SPR study data

- linear time trend of $0.14 - 0.26i/n$

from CUSUM Plot





Classification Terminology of Randomization Procedures

- pure ones
- with maximum tolerated imbalance (MTI)
- with final balance (FB)
- with maximum tolerated and final balance





- CR** complete randomization, tossing a fair coin, so the probability that patient i will receive treatment E is always $\frac{1}{2}$
- EBC(p)** Efron's biased coin, flip a biased coin $p = 2/3$ in favor of the less frequently allocated treatment
- UD(w, a, b)** (Wei's Urn Design) accounts adaptively for imbalance w, a, b .
- BSD(a)** big stick design, use CR allow for a *MTI* $a \in \{3, 4, 5\}$
- MP(a)** (Berger's Maximal Procedure) equiprobable version of big stick design
- Chen(p, a)** (Chen's Design) use EBC(p) allow for a *MTI* of $a \in \{2, 4\}$.
- RAR** random allocation rule, fix total sample size N and randomize so that half the patients receive treatment E , (*FB*)
- PBR(b)** permuted block randomization with block size $b \in \{2, 10\}$, implement RAR within each block (*MTI & FB*)
- ...etc.





several evaluation metrics are possible, averaged number of best guesses, balancing behavior, loss in estimation, etc.

ICH E9: The interpretation of statistical measures of uncertainty of the treatment effect and treatment comparisons should involve consideration of the potential **contribution of bias to the p-value**, confidence interval, or inference.

Assess the various randomization procedures with respect to

Metric: Level of Evidence

- averaged type 1 error probability over all sequences
- proportion of sequences which keep the 5% significance level





- Identify the “best practice” randomization procedure for the EnBand-Study by a comprehensive simulation study.
- Conduct a sensitivity study use η between 0.04, 0.08 and 0.16 as suitable values.

Decision Rule

- Select the design with the proportion of sequences with $\alpha \leq 0.05$ as close as possible to CR (2%)





... will use randomizeR, to conduct the evaluation and report the findings

current status of randomizeR

- implemented randomization procedures: CR, RAR, PBR, RPBR, HADA, MP, BSD, UD, TBD, EBC, GBC, CD, BBC
- ⇒ generating / saving a randomization sequence as .csv file
- implemented assessment criteria: selBias, chronBias, corGuess, imbal, setPower, combineBias
- ⇒ assessment and comparison of randomization procedures possible

in progress \ next steps

- assessment of linked criteria, randomization tests, time to event model, multiarm model
- bias corrected test
- development of a shiny app

ERDO Results of the Case Study



Randomization Procedure	Selection Bias	Linear-Time Trend Bias	Type I Error Probability [mean]	Type I Error Probability ≤ 0.05
CR	0.080	0.260	0.050	0.51
RAR	0.080	0.260	0.051	0.34
PBR(2)	0.080	0.260	0.073	0.00
PBR(10)	0.080	0.260	0.058	0.00
BSD(3)	0.080	0.260	0.052	0.10
BSD(4)	0.080	0.260	0.051	0.34
BSD(5)	0.080	0.260	0.050	0.45
MP(3)	0.080	0.260	0.055	0.00
MP(4)	0.080	0.260	0.053	0.01
MP(5)	0.080	0.260	0.052	0.06
EBC(2/3)	0.080	0.260	0.055	0.02
Chen(2)	0.080	0.260	0.060	0.00
Chen(4)	0.080	0.260	0.056	0.00
UD(0,1)	0.080	0.260	0.051	0.43
UD(1,2)	0.080	0.260	0.050	0.45

EP7 HEALTH 2013 - 602552





Table: Impact of selection bias and time trend on probability of type I error for different randomization procedures

Randomization Procedure	Selection Bias	Linear-Time Trend Bias	Type I Error Probability [mean]	Type I Error Probability ≤ 0.05
BSD(10)	0.080	0.260	0.050	0.53
BSD(15)	0.080	0.260	0.050	0.51
BSD(20)	0.080	0.260	0.050	0.50
BSD(25)	0.080	0.260	0.050	0.51
BSD(30)	0.080	0.260	0.050	0.52
BSD(35)	0.080	0.260	0.050	0.52
BSD(40)	0.080	0.260	0.050	0.53





Table: Impact of selection bias and time trend on probability of type I error for different randomization procedures

Randomization Procedure	Selection Bias	Linear-Time Trend Bias	Type I Error Probability [mean]	Type I Error Probability ≤ 0.05
UD(0,2)	0.080	0.260	0.051	0.43
UD(0,3)	0.080	0.260	0.051	0.43
UD(1,1)	0.080	0.260	0.050	0.46
UD(1,3)	0.080	0.260	0.050	0.44
UD(2,1)	0.080	0.260	0.050	0.47
UD(2,2)	0.080	0.260	0.050	0.46
UD(2,3)	0.080	0.260	0.050	0.46





Decision

- With a selection bias effect of $\eta = 0.08$ and a linear time trend of $0.14 - 0.26i/n$ it was shown, that the impact of the joint additive bias on the type I error probability inflation is kept to an acceptable minimum for the BSD(10). Acceptable minimum is given by at most 2% more or less sequences resulting in a type I error probability of 0.05 compared to CR. The comparison study includes CR, RAR, PBR(2,10), BSD(3,4,5,10,15,20,25,30,35,40), MP(4,5), EBC(2/3), Chen(2,4) and UD(0,1,2,1,2,3). Note, it results a loss in power of 1% by BSD(10).
- The sensitivity analysis taking into account values $\eta = 0.04$ to 0.12 by 0.04 and 0.13 to 0.39 by 0.13 for the slope confirms the findings.





- among other it is shown, that none of the randomization procedures perform uniformly best.
- practical restrictions, like balancing, risk of selection bias, risk of time trend bias may affect the choice of a randomization procedure.
- the choice the magnitude of η and θ have to be discussed within the practical context.
- at least a minimum effect (related to the clinical important effect size) should be assumed
- discussion of theses topics may help to understand the selection a randomization procedure within the particular/practical study settings and increase the level of evidence
- understand that the treatment effect could be hidden by bias, which may result from a randomization sequence





- software to do assessment is available, R package (*randomizeR*) (*Uschner, 2017*)
- start understanding effects with time to event data (*Rückbeil, 2017*)
- start understanding effects with multifactorial designs (*Uschner, 2017*)
- developed a uniform assessment criterion (*Schindler, 2016*)
- start understanding the effect of missing values on the test decision based on randomization based inference (*Heussen, 2016*)
- start understanding the randomization based inference in longitudinal linear mixed effects models (*Burger, 2017*)
- no yet completely developed a bias corrected test for all endpoints (*Kennes, 2015*)













- Ralf-Dieter Hilgers
- Christina Fitzner
- Nicole Heussen
- Lieven Kennes
- Simon Langer
- Martin Manolov
- Mui Pham
- Marcia Rückbeil
- David Schindler
- Antje Tasche
- Miriam Tamm
- Diane Uschner





-  Kennes, L. N., Cramer, E., Hilgers, R. E., and Heussen, N. (2011). The impact of selection bias on test decisions in randomized clinical trials *Statistics in Medicine* 2011; **30**:2573-2581.
-  Kennes, L. N. (2012). The impact of selection bias on test decisions in randomized clinical trials *Master Thesis Mathematics RWTH Aachen*
-  Kennes, L. N., Rosenberger William F., Hilgers, R.-D., (2015). Inference for blocked randomization under a selection bias model *Biometrics* 2015; **71**:y 979?984. doi.org/10.1111/biom.12334.
-  Langer S. The modified distribution of the t-test statistic under the influence of selection bias based on random allocation rule *Master Thesis, RWTH Aachen University, Germany*, 2014
-  Rückbeil M. The impact of selection bias on test decisions in survival analysis *Master Thesis, RWTH Aachen University, Germany*, 2015
-  Tasche A. Selection Bias bei mehr als zwei Behandlungsgruppen *Studienarbeit, RWTH Aachen University, Germany*, 2016
-  Tamm M, Cramer E, Kennes LN, Heussen N Influence of Selection Bias on the Test Decision - A Simulation Study *Methods of Information in Medicine* 2012; **51**:138-143. DOI: 10.3414/ME11-01-0043.
-  Tamm M, Hilgers RD. Chronological Bias in Randomized Clinical Trials Arising from Different Types of Unobserved Time Trends *Methods of Information in Medicine* 2014; **53**:501-510. DOI: 10.3414/ME14-01-0048.

