

Global interaction tests for predictive gene signatures in randomized clinical trials.

Test globali di interazione per profili genetici predittivi negli studi clinici randomizzati.

Federico Rotolo and Stefan Michiels

Abstract With the advent of the targeted therapy era in oncology, prognostic and predictive gene signatures are becoming increasingly important in clinical research and even in clinical practice. We investigate an approach to develop a gene signature as treatment modifier in a phase III clinical trial. We propose to apply first a permutation procedure in a survival model that controls the family-wise error rate at a pre-specified level. Only if the global test is significant, a classifier can be developed for predicting the treatment effects for future patients. We present the results of a simulation study with 100 candidate biomarkers in a relatively small clinical trial and compare different permutation tests to control the type I error.

Sommario Con l'avvento delle terapie mirate in oncologia, i profili genici prognostici e predittivi stanno diventando sempre più importanti sia nella ricerca che addirittura nella pratica clinica. Presentiamo un approccio per sviluppare profili genetici che modificano l'effetto del trattamento. Proponiamo qui di applicare per prima cosa una procedura di permutazione in un modello di durata che controlli il tasso di errore globale ad un livello fissato. Solo se il test globale è significativo, si può sviluppare un classificatore per predire gli effetti del trattamento per i pazienti futuri. Presentiamo uno studio di simulazione con 100 marcatori candidati in uno studio clinico relativamente piccolo e compariamo diversi test di permutazione per controllare l'errore di tipo I.

Key words: Gene signatures, global tests, interaction, family-wise error, predictive marker, permutation tests, penalization, AIC.

Federico Rotolo

Gustave Roussy, Service de Biostatistique et d'Épidémiologie, F-94805 Villejuif, France
e-mail: federico.rotolo@gustaveroussy.fr

Stefan Michiels

Gustave Roussy, Service de Biostatistique et d'Épidémiologie, F-94805 Villejuif, France.
Université Paris-Sud, F-94805 Villejuif, France
e-mail: stafan.michiels@gustaveroussy.fr.

1 Introduction

With the advent of the targeted therapy era, molecular signatures have an increasing role in anticipating the prognosis of patients (*prognostic* biomarkers) or predicting how they respond to specific treatments (*predictive* biomarkers or *effect modifiers*) [11, 2]. After the assessment of the analytical and the clinical validity of a gene signature in several independent patient cohorts, one must consider to design a clinical trial to establish its clinical utility [4, 13], which can be defined as

evidence of improved measurable clinical outcomes, and its usefulness and added value to patient management decision-making compared with current management without genetic testing

accordingly to the Evaluation of Genomic Applications in Practice and Prevention Initiative [13]. Gene signatures must be tested within a clinical trial to evaluate their utility as prognostic factors or their predictiveness as treatment modifiers, i. e. their ability to discriminate patients who will benefit from the treatment from the others.

More and more frequently, in phase III randomized clinical trials several biomarkers are tested as candidate predictors of the treatment effect. Of note, as a gene signature is not known beforehand, these tests are in the context of exploratory analyses, unless the study has been designed in a 2-stage adaptive way [3]. Recently, Michiels *et al.* proposed a permutation approach for global interaction tests (GITs) to control the type I error when there is correlation among the biomarkers [7]; we review these methods in Section 2. With next-generation sequencing techniques provide panels of at least 100 cancer genes in retrospective analyses of clinical trials are available. In Section 3, we extend the previously proposed permutation approach to a higher dimensional setting by exploring the added value of a ridge regression and a variable selection using Akaike's Information Criterion (AIC [1]). In the present study we consider only DNA sequencing of mutations (e. g. exome sequencing), so we restrict ourselves to the case of binary biomarkers.

2 Global interaction tests for gene signatures as treatment modifiers

In all-comers clinical trials, many candidate biomarkers (k) are often tested as possible treatment-effect modifiers. Rothwell *et al.* [10] put forward that the only reliable statistical approach for assessing the predictiveness of biomarkers is to test for interaction between a subgroup and the treatment effect acknowledging the need for pre-specification of subgroups. We assume a time-to-event outcome and that the test for interaction is for a multiplicative effect on a relative scale. In order to control the family-wise type I error, accounting for dependence structures, Michiels *et al.* [7] proposed five different permutation tests, using single-biomarker statistics and composite statistics based on treatment-by-biomarker interactions in a Weibull accelerated failure time (AFT) model [8, 5]. We consider here three of these tests:

Max-Single Wald (MSW), Sum-Single Wald (SSW), and Composite Difference (CD). The first two tests (MSW, SSW) are based on k biomarker-specific models, each containing three factors: the treatment, the biomarker, and their interaction. The Wald statistics for the interactions in the biomarker-specific models are computed and their maximum (MSW) or their sum (SSW) across the biomarker-specific models is taken. On the other hand, the idea behind the CD test is to define a scaled linear combination of all the k biomarkers — the composite biomarker score — combining the individual interaction signals. It is based on the parameters of the interaction effects of a full model with $2k + 1$ factors: the treatment, the k biomarkers, and the k interactions. The CD statistic is the absolute difference in two concordance probabilities, the concordance probability that a randomly chosen control patient will outlive a treated patient for positive biomarker minus the same concordance probability for negative biomarker score. Permutations rearrange the patients only within each treatment group. For the permutation tests to be valid the data does not need to follow a Weibull distribution.

Michiels *et al.* [7] showed that the three tests provided an appropriate control of the global type I error under the null scenarios, but the CD test was slightly too conservative when there was just one prognostic biomarker. Under the alternative scenarios, the MSW test was optimal, followed by the SSW, when there was one truly treatment-modifying biomarker. The SSW or CD statistics were optimal when there were two or three independent treatment-modifying biomarkers.

Based on the permutation GITs, a 2-step strategy for developing a signature from many candidate treatment-effect modifying biomarkers in a phase III trial can be devised. First, a GIT (either MSW, SSW, or CD) is performed to control the type I error at a pre-specified level. Only if the GIT is significant, a gene signature can be developed in a second step; to this end, AFT models can be used and a resampling procedure such as cross-validation could be employed to estimate treatment effects in biomarker-defined patient subsets [6]. The application of the AFT model to the full development data provides a classifier for future patients [12].

3 Extension of the global interaction tests to a higher dimensional setting

With next-generation sequencing techniques, the number of candidate biomarkers is usually very high, dangerously close or even greater than the number of events typically observed in clinical trials. The high number of interactions to test can compromise the performances of the GITs or even make them infeasible, notably those based on full models (CD). In such scenarios, reduction of the dimensionality of the problem is needed. We considered two solutions: penalization and preselection.

A huge literature exists on penalized regression for time-to-event data; the simplest solutions consist in models with L_2 (ridge) [15] or L_1 (least absolute shrinkage and selection operator, LASSO) [14] penalties, but more complex extensions are available [16, 17]. Since our primary interest was the control of global type I error

and not the variable selection *per se*, we explored the added value of ridge regression when fitting a Weibull AFT model with $2k + 1$ variables before calculating the CD test (denoted as $L_2 + CD$). The L_2 penalties were chosen by minimum AIC over a grid of plausible values.

The other way that we considered for simplifying the problem was computing the GITs after a preselection of the most relevant biomarkers. We adopted a stepwise model selection based on the minimum AIC, starting from the model containing only the treatment effect. The effects of the biomarker and of their interactions with the treatment were added or dropped iteratively, under three constraints: the treatment effect could not be removed; each biomarker effect could be removed provided that its interaction with the treatment was not in; the interaction of a biomarker with the treatment could be added provided that its main effect was in. In the final chosen model, we calculated the GITs using the parameter estimates of the retained interaction effects and their estimated variances (AIC+SSW, AIC+MSW, and AIC+CD).

Table 1 Description of the 5 treatment by true biomarker scenarios used in simulation study. **B-**: active Biomarker-negative; **B+**: active Biomarker-positive; $\rho = 1$: Exponential times; $\rho = 1.25$: Weibull times.

Scenario	Median survival time (years)				Average censoring probability			
	Control		Treatment		Control		Treatment	
	B-	B+	B-	B+	$\rho = 1$	$\rho = 1.25$	$\rho = 1$	$\rho = 1.25$
(1) Complete null	1.0	1.0	1.0	1.0	0.10	0.06	0.11	0.06
(2) One prognostic active marker, strong effect size	0.6	1.2	0.9	1.8	0.09	0.05	0.18	0.13
(3) One treatment-modifying active biomarker, strong effect size	0.8	0.8	0.8	1.6	0.06	0.03	0.15	0.10
(4) Five treatment-modifying active biomarkers, medium effect size	0.8	0.8	0.8	1.2	0.06	0.03	0.35	0.32
(5) Five treatment-modifying active biomarkers, medium effect size, correlation = 0.5 by 10	0.8	0.8	0.8	1.2	0.06	0.03	0.35	0.32

We performed a simulation study to investigate, in the presence of 100 candidate biomarkers, the performance of the GITs and the impact of penalization or preselection. We considered a metastatic setting with median survival time (mst) of about one year. We first present the case of $n = 300$ patients, 150 per treatment arm, then we double the sample size to $n = 600$ patients, 300 per arm. We generated survival times via exponential ($\rho = 1$) and then with Weibull ($\rho = 1.25$, i. e. increasing risk) random variables. Assuming 3-year uniform accrual and 2-year follow-up, we generated independent censoring as uniform times between 2 and 5 years.

We simulated 100 binary biomarkers with prevalence of $1/2$ by truncation of normal multivariate random variables. Biomarkers were independent of each other in all but one scenario, in which correlation of the underlying normal random variables was specified for 10-biomarker blocks. Table 1 describes the five scenarios. In the complete null scenario, neither the treatment nor the biomarkers had any effect. In the second null scenario, 99 biomarkers had no effect as in the complete null scenario, whereas one strong prognostic biomarker doubled the mst and the treatment increased by 50% the mst; none of the biomarkers interacted with the treatment. In the alternative scenario 3 there was one activating biomarker, i.e. the treatment had no effect except for patients with one specific biomarker positive, whose mst was doubled. The fourth scenario was similar to the third, but with five activating biomarkers, the benefit of each being smaller (+50% of mst). The last scenario was as the fourth, with the addition of correlation (0.5 for the underlying normal variables) within random blocks of 10 biomarkers. The censoring proportions (last four columns of Table 1) ranged between 3 and 35%.

We evaluated the GITs in terms of their empirical rejection probabilities and p-values. We simulated 250 trials per scenario and we computed the GITs in 500 independent permutations within each treatment arm, without moving any patient from an arm to the other. We estimated the p-values as the proportion of permutations for which the test statistic was greater than for the data prior to permutation. Strictly speaking, this approach tests the broader hypothesis of neither main nor interaction effects, but it is particularly sensitive to the presence of interaction effects; indeed, permutation tests are not able to test for interaction only [9].

We provide the power estimates at a significance level $\alpha = 0.05$, that is the proportion of the 250 datasets for which the null hypothesis is rejected at a 5% level (Tab. 2-3 for $\rho = 1$ and $\rho = 1.25$, respectively; Fig. 1). Of note, if the rejection probability of a test is actually 0.05, then the exact 95% confidence interval for its empirical rejection probability over 250 repetitions is (0.024, 0.080). However, permutation tests can have biased power estimates [9]. Hence, we computed unbiased estimates of the mean p-values as the average of the p-values across the 250 replications (Tab 4-5; Fig. 2).

Under the complete null scenario 1, the empirical power of all the GITs was almost always within the exact 95% confidence interval computed for an actual rejection probability of 0.05; the mean p-values were all close to 0.50.

In the presence of one prognostic biomarker (null scenario 2) the CD test was more conservative, notably with $n = 600$, with rejection probabilities less than 0.02 and mean p-values between 0.60 and 0.70; neither penalization nor preselection (L_2 +CD or AIC+CD) could alleviate this issue appreciably. The SSW and AIC+MSW tests also had empirical rejection probabilities significantly different from 0.05 with $n = 600$ and $\rho = 1$, but their mean p-values were reassuring.

In the three alternative scenarios, the power of the GITs was generally quite low with $n = 300$ and $\rho = 1$ (3 to 40%); the mean p-values were greater or equal to 0.17 and reached or exceeded 0.50 sometimes. When the sample size was higher ($n = 600$) or the risk increased along time ($\rho = 1.25$), the power of all the GITs was sensibly higher, up to 0.93, with mean p-values ranging from 0.02 to 0.52. In

Table 2 Results of the simulation study with $n=300$ and $n=600$, and $\rho = 1$. Rejection probabilities for the 5 scenarios. **SSW**: Sum Single Wald statistic; **MSW**: Maximum Single Wald statistic, **CD**: Composite Difference statistic, **L₂+CD**: Composite Difference statistic in an L2-penalised model (ridge), **AIC+SSW**: Sum Single Wald statistic after AIC selection, **AIC+MSW**: Maximum Single Wald statistic after AIC selection, **AIC+CD**: Composite Difference statistic after AIC selection

$\rho = 1$			Test						
<i>Rejection probabilities</i>						L_2 + CD	AIC + SSW	AIC + MSW	AIC + CD
Scenario	1	n=300	0.06	0.03	0.05	0.06	0.06	0.05	0.04
		n=600	0.05	0.05	0.07	0.07	0.04	0.03	0.07
	2	n=300	0.05	0.05	0.01	0.00	0.03	0.05	0.01
		n=600	0.02	0.05	0.01	0.01	0.04	0.02	0.02
	3	n=300	0.12	0.30	0.07	0.05	0.09	0.23	0.23
		n=600	0.28	0.69	0.14	0.28	0.38	0.75	0.69
	4	n=300	0.16	0.14	0.09	0.03	0.14	0.16	0.13
		n=600	0.44	0.42	0.18	0.29	0.50	0.47	0.36
	5	n=300	0.40	0.31	0.07	0.17	0.11	0.16	0.24
		n=600	0.80	0.70	0.25	0.57	0.48	0.48	0.51

Table 3 Results of the simulation study with $n=300$ and $n=600$, and $\rho = 1.25$. Rejection probabilities for the 5 scenarios. **SSW**: Sum Single Wald statistic; **MSW**: Maximum Single Wald statistic, **CD**: Composite Difference statistic, **L₂+CD**: Composite Difference statistic in an L2-penalised model (ridge), **AIC+SSW**: Sum Single Wald statistic after AIC selection, **AIC+MSW**: Maximum Single Wald statistic after AIC selection, **AIC+CD**: Composite Difference statistic after AIC selection

$\rho = 1.25$			Test						
<i>Rejection probabilities</i>						L_2 +	AIC +	AIC +	AIC +
			SSW	MSW	CD	CD	SSW	MSW	CD
Scenario	1	n=300	0.05	0.04	0.07	0.04	0.08	0.07	0.04
		n=600	0.06	0.05	0.09	0.07	0.02	0.04	0.06
	2	n=300	0.06	0.04	0.00	0.00	0.06	0.04	0.02
		n=600	0.06	0.03	0.00	0.00	0.04	0.04	0.03
	3	n=300	0.26	0.60	0.10	0.07	0.12	0.42	0.41
		n=600	0.48	0.93	0.23	0.39	0.68	0.94	0.84
	4	n=300	0.30	0.31	0.09	0.10	0.19	0.22	0.26
		n=600	0.64	0.62	0.39	0.49	0.78	0.74	0.68
	5	n=300	0.61	0.50	0.12	0.20	0.18	0.22	0.34
		n=600	0.90	0.84	0.48	0.76	0.74	0.72	0.74

Table 4 Results of the simulation study with $n=300$ and $n=600$, and $\rho = 1$. Unbiased estimate of mean p-value for the 5 scenarios. **SSW**: Sum Single Wald statistic; **MSW**: Maximum Single Wald statistic, **CD**: Composite Difference statistic, **L₂+CD**: Composite Difference statistic in an L2-penalised model (ridge), **AIC+SSW**: Sum Single Wald statistic after AIC selection, **AIC+MSW**: Maximum Single Wald statistic after AIC selection, **AIC+CD**: Composite Difference statistic after AIC selection

$\rho = 1$			Test						
<i>Mean p-values</i>						L_2 + CD	AIC + SSW	AIC + MSW	AIC + CD
Scenario	1	n=300	0.50	0.49	0.49	0.47	0.51	0.52	0.48
		n=600	0.48	0.53	0.48	0.50	0.51	0.52	0.48
	2	n=300	0.51	0.53	0.66	0.64	0.53	0.54	0.63
		n=600	0.52	0.54	0.70	0.70	0.55	0.56	0.64
	3	n=300	0.39	0.28	0.50	0.50	0.37	0.31	0.32
		n=600	0.22	0.10	0.35	0.24	0.15	0.07	0.08
	4	n=300	0.32	0.36	0.49	0.52	0.35	0.36	0.36
		n=600	0.15	0.18	0.29	0.22	0.13	0.15	0.17
	5	n=300	0.17	0.24	0.45	0.38	0.33	0.31	0.27
		n=600	0.05	0.06	0.25	0.13	0.11	0.12	0.11

Table 5 Results of the simulation study with $n=300$ and $n=600$, and $\rho = 1.25$. Unbiased estimate of mean p-value for the 5 scenarios. **SSW**: Sum Single Wald statistic; **MSW**: Maximum Single Wald statistic, **CD**: Composite Difference statistic, **L₂+CD**: Composite Difference statistic in an L2-penalised model (ridge), **AIC+SSW**: Sum Single Wald statistic after AIC selection, **AIC+MSW**: Maximum Single Wald statistic after AIC selection, **AIC+CD**: Composite Difference statistic after AIC selection

$\rho = 1.25$			Test						
<i>Mean p-values</i>						L_2 + CD	AIC + SSW	AIC + MSW	AIC + CD
Scenario	1	n=300	0.52	0.52	0.51	0.51	0.48	0.49	0.49
		n=600	0.48	0.52	0.48	0.49	0.50	0.51	0.51
	2	n=300	0.48	0.51	0.68	0.71	0.57	0.57	0.68
		n=600	0.48	0.47	0.78	0.74	0.54	0.53	0.62
	3	n=300	0.27	0.15	0.46	0.52	0.31	0.21	0.20
		n=600	0.13	0.02	0.31	0.19	0.06	0.02	0.03
	4	n=300	0.21	0.25	0.38	0.49	0.28	0.26	0.23
		n=600	0.08	0.09	0.17	0.12	0.05	0.06	0.07
	5	n=300	0.10	0.15	0.42	0.39	0.24	0.23	0.17
		n=600	0.02	0.03	0.16	0.06	0.05	0.06	0.06

the presence of only one strongly predictive biomarker (scenario 3), the MSW and the AIC+MSW tests performed the best, but also the AIC+CD test had competitive power. The SSW and MSW tests were the best in the case of five predictive biomarkers (scenario 4), together with their preselection counterparts AIC+SSW and AIC+MSW; these latter were appreciably better with large datasets ($n = 600$). The CD test had generally poor power in alternative scenarios as compared to the other tests, with mean p-values that were often the highest. Even though penalization seemed able to improve its performances for the larger sample size, minimum AIC stepwise preselection increased substantially its power.

AIC preselection was able to improve the power of the SSW and MSW tests provided that enough data was at hand ($n = 600$). This result suggests that in the case of few observations and many candidate biomarkers ($n = 300$), stepwise selection does not select the right biomarkers, yielding to possibly wrong conclusions. Nevertheless, results in the last scenario (5) suggest that a strong correlation pattern between active and inactive biomarkers can compromise the correctness of preselection.

Globally, in the alternative scenarios doubling the sample size ($n = 600$) made the mean p-values of all the tests about 2.6-fold smaller and their power was on average 3.4 times higher than for $n = 300$.

4 Conclusion

Clinical trials are more and more often designed for testing gene signatures as treatment effect modifiers. A safe gatekeeping approach consists in two steps: first, a GIT is run to control the overall type I error at a prespecified level; only if the GIT is significant, the selection of the most plausible effect modifiers is undertaken. If many biomarkers are expected to be truly associated with differential treatment efficacy, the SSW or MSW tests may be the preferable choice. In the present paper we studied the tests with a high number of biomarkers (100) and we explored the added value of ridge penalization and of preliminary AIC variable selection. We showed that AIC preselection can improve the performance of these GITs, while ridge penalization did not show any advantage, globally. Present research focuses on improving the power of GITs, on extending them to the case of continuous biomarkers, as in RNA sequencing (gene expression), and to develop approaches that combine the inference step and the variable selection step.

References

1. Akaike, H.: A new look at the statistical model identification. *IEEE T Automat Contr* **19**(6), 716–723 (1974). ↩
2. Arnedos, M., André, F., Farace, F., Lacroix, L., Besse, B., Robert, C., Soria, J.C., Eggermont, A.M.: The challenge to bring personalized cancer medicine from clinical trials into routine clinical practice: the case of the Institut Gustave Roussy. *Mol Oncol* **6**(2), 204–210 (2012). ↩

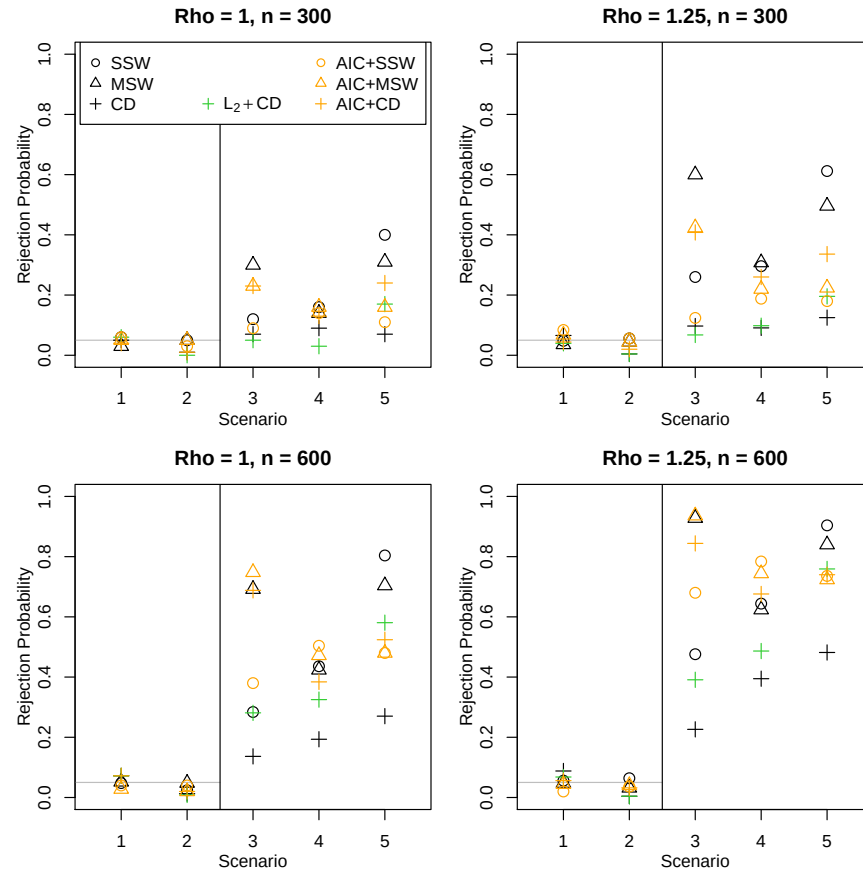


Fig. 1 Empirical rejection probabilities of the GITs with Exponential ($\rho = 1$, left) and Weibull ($\rho = 1.25$, right) event times, in 250 datasets of size 300 (top) and 600 (bottom).

3. Buyse, M., Michiels, S.: Omics-based clinical trial designs. *Curr Opin Oncol* **25**(3), 289–95 (2013). [↔](#)
4. Haga, S.B., Khoury, M.J., Burke, W.: Genomic profiling to promote a healthy lifestyle: not ready for prime time. *Nat Genet* **34**(4), 347–50 (2003). [↔](#)
5. Klein, J.P., Moeschberger, M.L.: Survival analysis: techniques for censored and truncated data, 2nd edn. Springer Verlag (2003). [↔](#)
6. Matsui, S., Simon, R., Qu, P., Shaughnessy, J.D., Barlogie, B., Crowley, J.: Developing and validating continuous genomic signatures in randomized clinical trials for predictive medicine. *Clin Cancer Res* **18**(21), 6065–73 (2012). [↔](#)
7. Michiels, S., Potthoff, R.F., George, S.L.: Multiple testing of treatment-effect-modifying biomarkers in a randomized clinical trial with a survival endpoint. *Stat Med* **30**(13), 1502–18 (2011). [↔](#)
8. Odell, P.M., Anderson, K.M., D’Agostino, R.B.: Maximum likelihood estimation for interval-censored data using a weibull-based accelerated failure time model. *Biometrics* pp. 951–9 (1992)

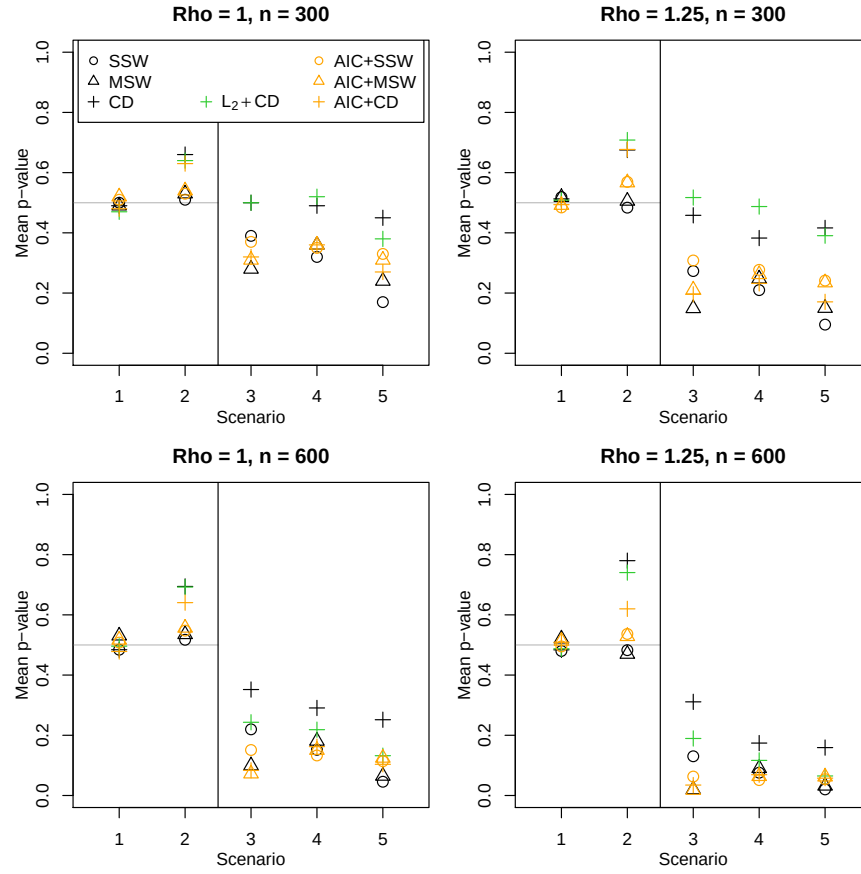


Fig. 2 Mean p-values of the GITs with Exponential ($\rho = 1$, left) and Weibull ($\rho = 1.25$, right) event times, in 250 datasets of size 300 (top) and 600 (bottom).

9. Potthoff, R.F., Peterson, B.L., George, S.L.: Detecting treatment-by-centre interaction in multi-centre clinical trials. *Stat Med* **20**(2), 193–213 (2001). [↔](#)
10. Rothwell, P.M.: Subgroup analysis in randomised controlled trials: importance, indications, and interpretation. *Lancet* **365**(9454), 176–86 (2005). [↔](#)
11. Sawyers, C.: Targeted cancer therapy. *Nature* **432**(7015), 294–7 (2004). [↔](#)
12. Simon, R.: Clinical trials for predictive medicine. *Stat Med* **31**(25), 3031–40 (2012). [↔](#)
13. Teutsch, S.M., Bradley, L.A., Palomaki, G.E., Haddow, J.E., Piper, M., Calonge, N., Dotson, W.D., Douglas, M.P., Berg, A.O.: The evaluation of genomic applications in practice and prevention (EGAPP) initiative: methods of the EGAPP working group. *Genet Med* **11**(1), 3–14 (2009). [↔](#)
14. Tibshirani, R., et al.: The lasso method for variable selection in the Cox model. *Stat Med* **16**(4), 385–95 (1997). [↔](#)
15. Verweij, P.J., Van Houwelingen, H.C.: Penalized likelihood in Cox regression. *Stat Med* **13**(23–24), 2427–36 (1994). [↔](#)
16. Zhang, H.H., Lu, W.: Adaptive lasso for Cox’s proportional hazards model. *Biometrika* **94**(3), 691–703 (2007). [↔](#)

17. Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. *J Roy Stat Soc B Met* **67**(2), 301–20 (2005). $\hookrightarrow$