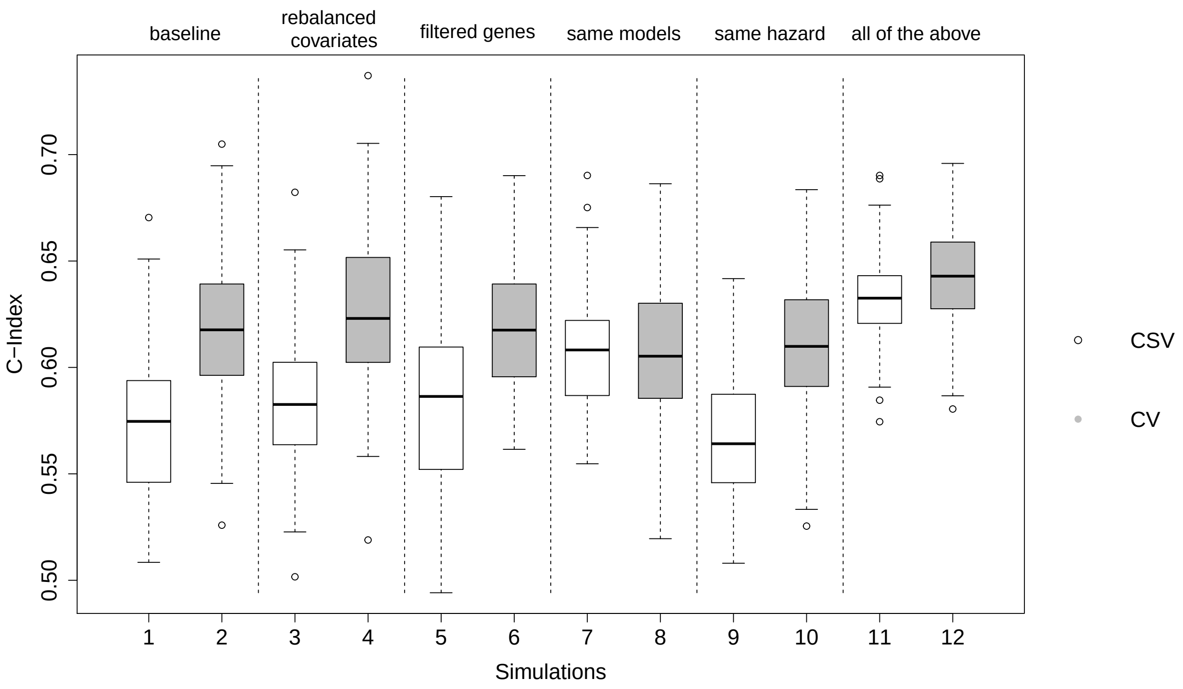
Supplementary Materials for ”Dataset heterogeneity in the validation of prediction models across studies”

Yuqing Zhang, Christoph Bernau, Giovanni Parmigiani, Levi Waldron

Contents

1 Boxplots of CV and CSV scores

Breast Cancer / Ridge Regression



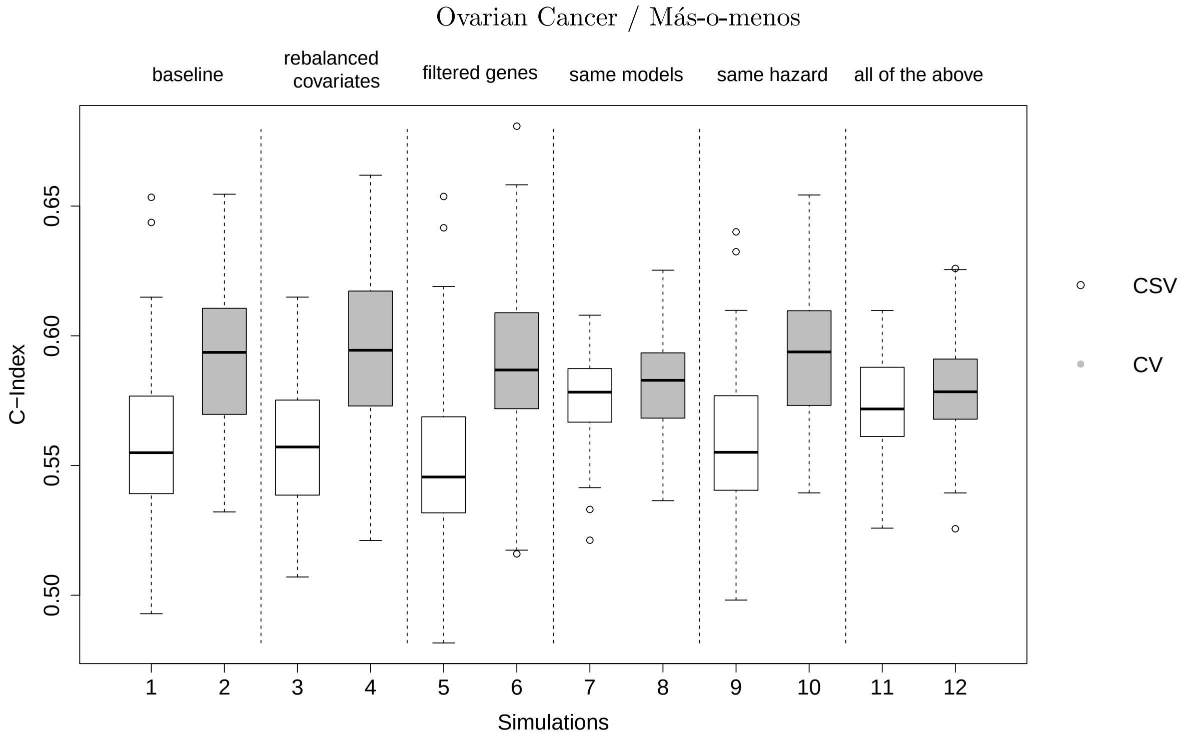


Figure 1: Simulation results comparing performances of cross validation and cross-study validation for combinations breast cancer / ridge regression, and ovarian cancer / Más-o-menos . Panel names are the same as Figure 2 in the main paper. Results for this two other dataset / algorithm combinations accord with the trends in Figure 2 and our conclusions.

2 Availability of Covariates in Breast Cancer Data

name		CAL	MNZ	MSK	TAM1	TAM2	TRB	UNT	VDX	Overall
ER+		75	162	57	507	325	134	86	209	1555
pgr	availablity	75	/	57	65	321	/	52	/	570
	0	17	/	15	5	72	/	1	/	110
	1	58	/	42	60	249	/	51	/	460
her2	availablity	/	/	49	/	149	/	/	/	198
	0	/	/	0	/	110	/	/	/	110
	1	/	/	49	/	39	/	/	/	88
size	availablity	74	162	57	341	164	134	86	209	1227
	0	29	96	8	138	77	76	55	203	698
	1	45	66	49	203	87	58	31	6	529
node	availablity	75	162	57	493	164	134	86	209	1380
	0	32	/	15	331	70	/	/	/	1039
	1	43	/	42	162	94	/	/	/	341
age	availablity	74	162	57	371	324	134	86	209	1417
	0	31	38	22	32	59	95	32	90	399
	1	43	124	35	339	265	39	54	119	1018
grade	availablity	74	162	/	346	286	132	72	142	1214
	1	5	27	/	77	56	29	29	4	227
	2	26	119	/	202	134	68	31	36	616
	3	43	16	/	67	96	35	12	102	371

Table 1: Availability of covariates in 8 breast cancer data sets. The table reports the number of patients left in the datasets after initial cleaning. Patient age and tumor size are dichotomized. size = tumor size, age = age of patients at diagnosis, node = nodal status, grade = historical grade. Datasets acronyms are the same as Table 1 in the main article.

3 Hazard Ratio of Covariates

Cov	HR	Confidence Interval	#datasets
Size	1.96	[1.46 , 2.62]	7
Age	1.2	[0.86 , 1.65]	6
Histological Grade	1.65	[1.24 , 2.19]	7

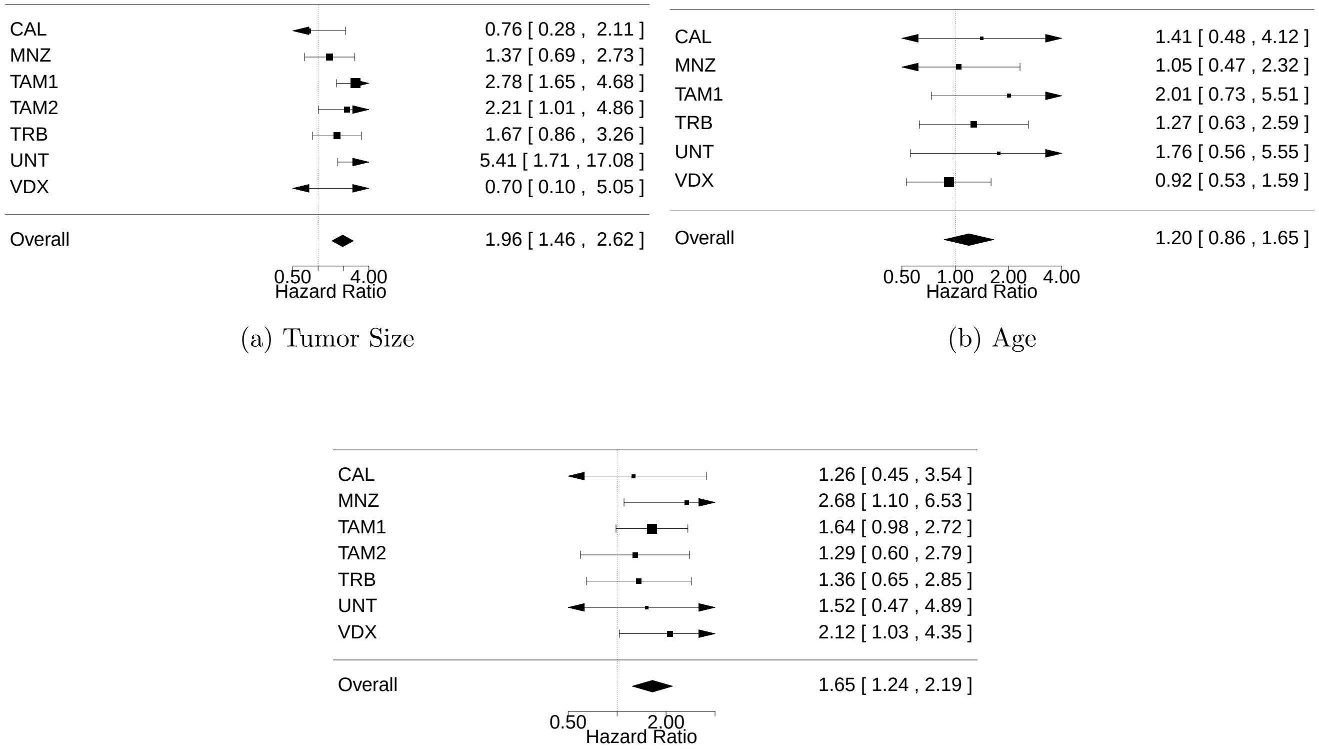
Table 2: **Synthesized hazard ratios of all covariates in the breast cancer datasets.** The three-level histological grade is dichotomized for the calculation of HR, by combining the low and intermediate levels as 0, and treating the high level as 1. Row labels: Cov = covariate names, HR = hazard ratio synthesized across all base datasets by fixed-effects meta-analysis. Interval = 95% Confidence Interval, #datasets = Number of studies available for hazard ratio. Tumor size and histological grade, but not age dichotomized at 50 years, are significantly associated with disease and metastasis-free survival.

Cov	HR	Confidence Interval	#datasets
Age	1.84	[1.51 , 2.24]	5
Debulking	1.48	[1.23 , 1.78]	5

Table 3: **Synthesized hazard ratios of all covariates in the ovarian cancer datasets.** Row labels: Cov = covariate names, HR = hazard ratio synthesized across all datasets by fixed-effects meta-analysis. Interval = 95% Confidence Interval, #datasets = Number of studies available for hazard ratio. All factors are significantly associated with survival.

4 Forest Plots of Hazard Ratio of the Covariates

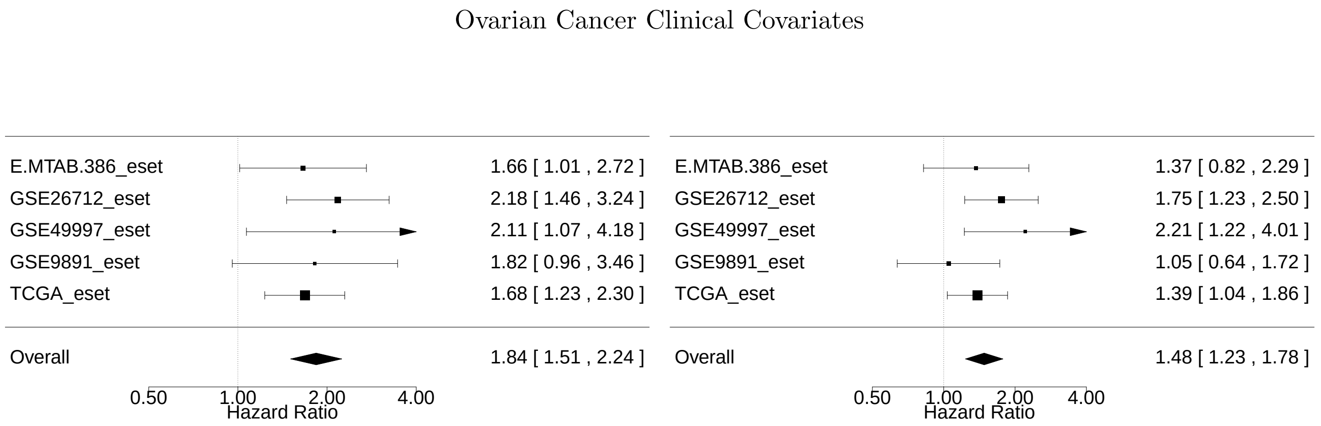
Breast Cancer Clinical Covariates



(a) Tumor Size

(b) Age

(c) Histological Grade



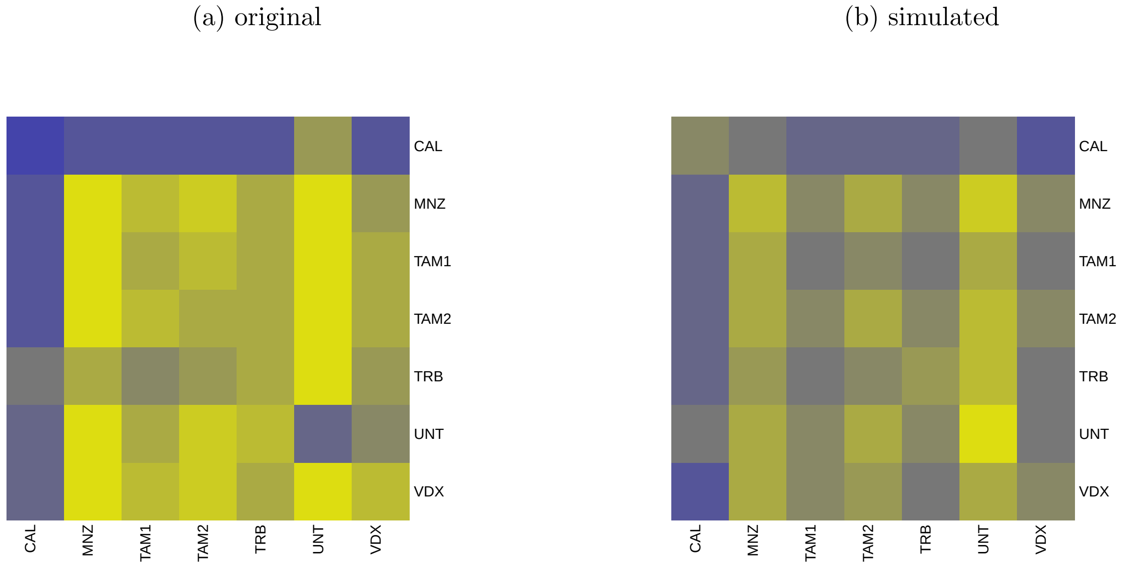
Ovarian Cancer Clinical Covariates

(a) Age

(b) Debulking

Figure 2: Forest plots of hazard ratios to illustrate the relative strength of covariate impact in the studies. For breast cancer, at a certain level of confidence, the impacts of all these covariates significantly differ from no effect on the corresponding survival time, based on the deviance of overall confidence interval from the null hypothesis. For ovarian cancer, both factors significantly associate with survival. To show that the covariates we consider are prognostic further explains our finding that the distribution of these variables do not influence the validation performance.

5 Cross-study validation matrices using Más-o-menos on the original and simulated breast cancer datasets

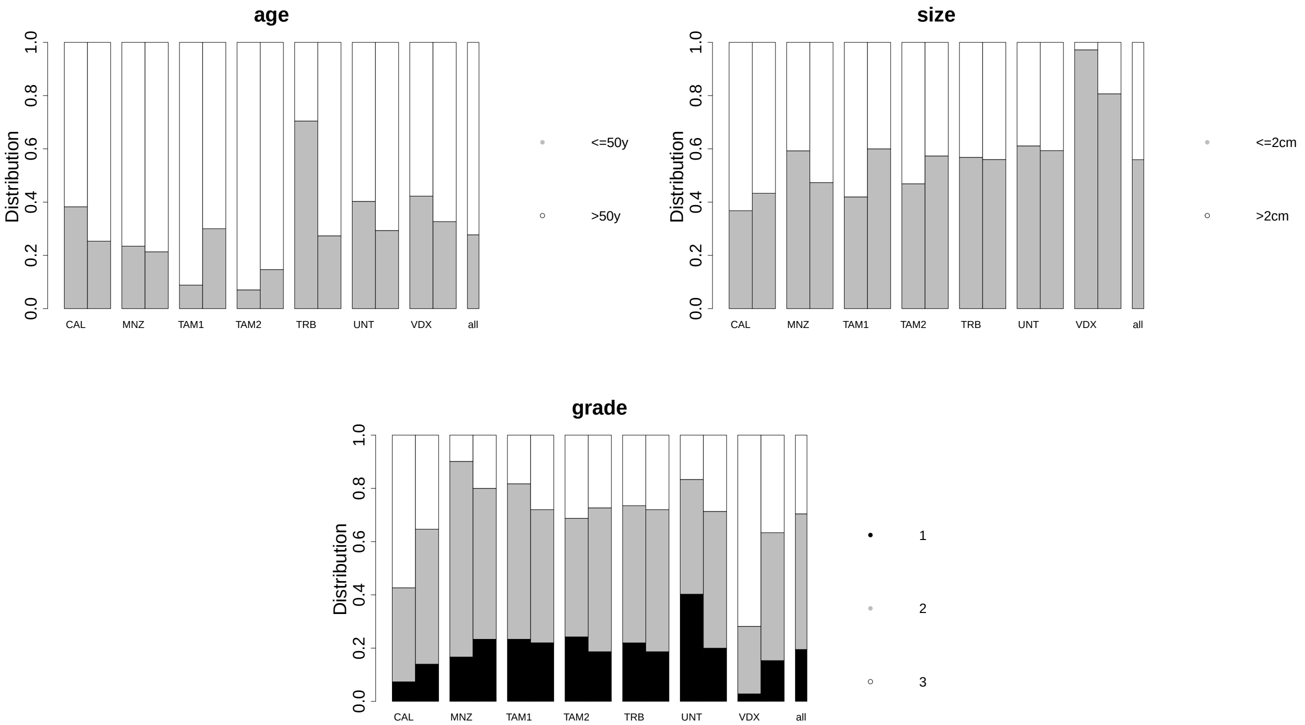

(a) original

(b) simulated

Figure 3: Cross-study validation matrices using Más-o-menos algorithm on ER+ breast cancer data, showing comparability of simulated and experimental data for the purpose of discrimination of disease and metastasis-free survival. Panel (a) displays C-indices for training and validation on each pair of original datasets, with the diagonal showing results of 4-fold CV. Panel (b) displays the equivalent heatmap, of the Z matrix averaged over 100 simulations, using simulated data generated by non-parametric bootstrap resampling of individuals and parametric simulation of censored time to event outcomes based on models estimated from the original base sets. Within and across-study validation performance is similar between original and simulated data, with the exception of the outlier CAL dataset, which shows a larger difference in CV-CSV performances than the other datasets. Note that Bernau *et al.* (2014) displayed one additional outlier set that was eliminated from this study due to unavailability of histological grade.

6 Bar-plots of covariate distributions before and after balancing

Breast Cancer



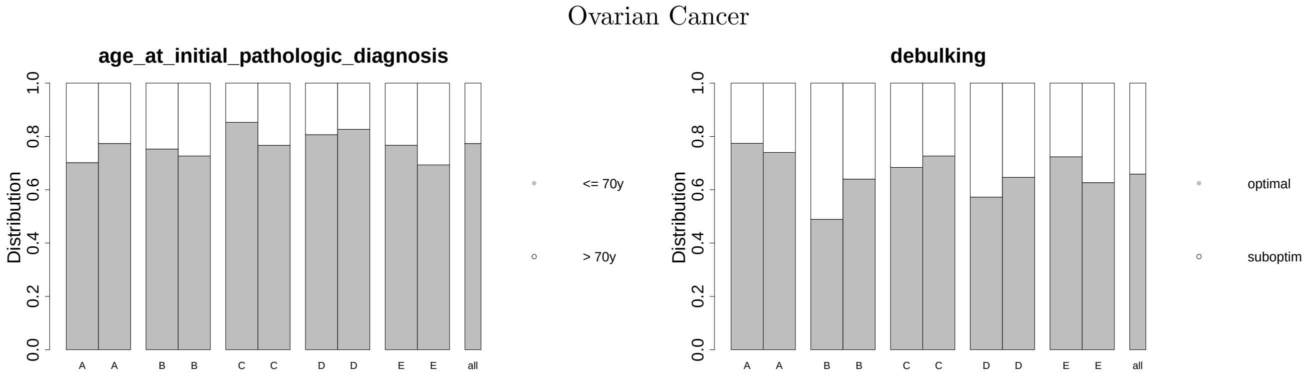


Figure 4: Comparison before and after balancing covariates. It is to illustrate the balancing effect after we adjust the probability of re-sampling. The single bar on the most right shows the overall distribution of the covariate across all original data sets, which we expect to be the distribution of this covariate in each simulated set after balancing. From left to right, every two close together bars represent the distribution in one data set before(the left bar) and after(the right bar) balancing. Changing the probability will not make the proportion of levels after balancing equal to the overall distribution, so the bars are not exactly the same. But we can observe that the distributions after balancing are closer to the overall distribution than those without balancing. Severe bias is adjusted such as tumor size in VDX and age in TRANSBIG. The figures prove that we actually balanced the prevalence of covariates as we wanted. As a back up to the simulation, it enhances our conclusion about the irrelevance of covariate distributions to C-index performances.

7 Spaghetti Plots

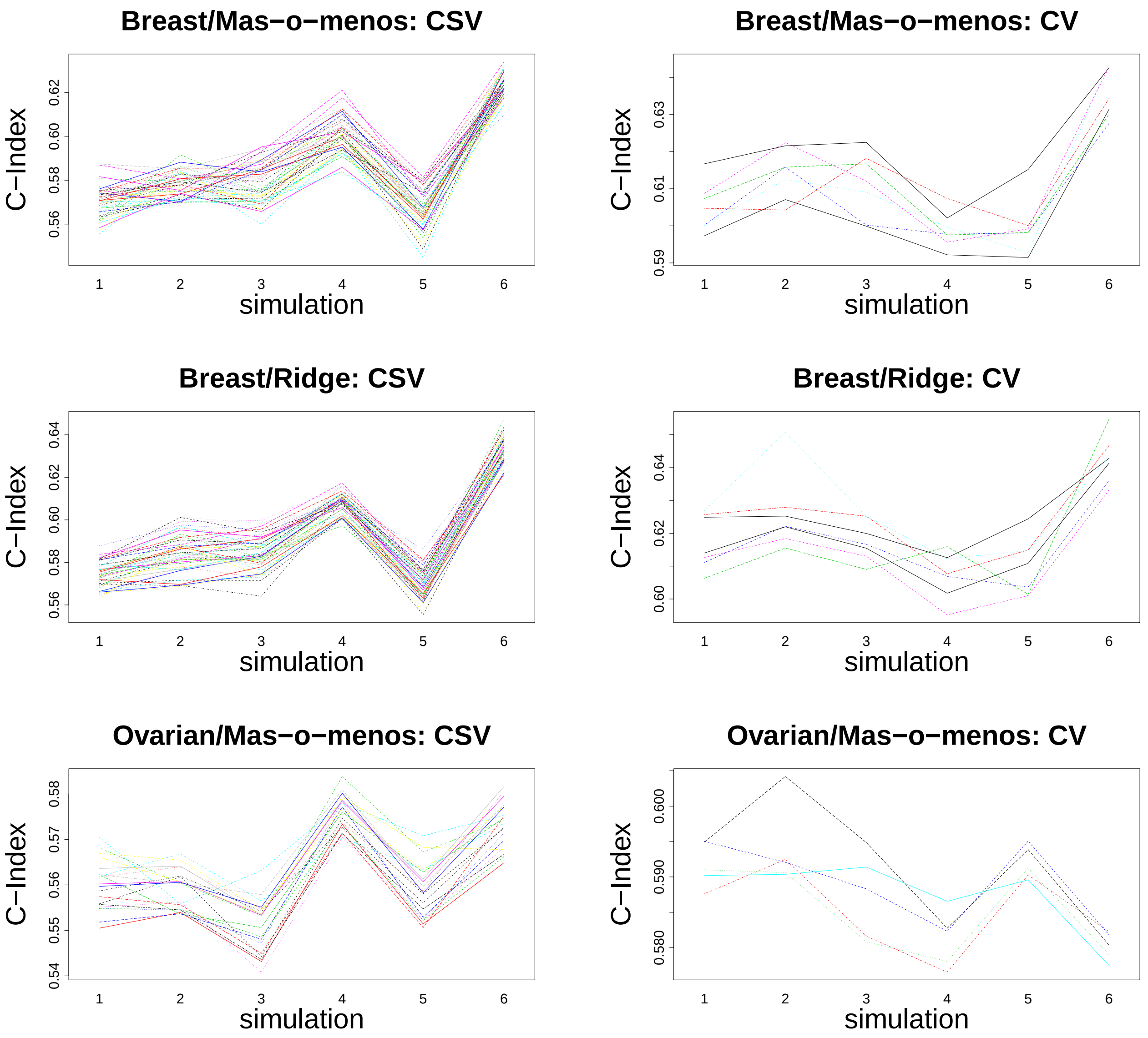
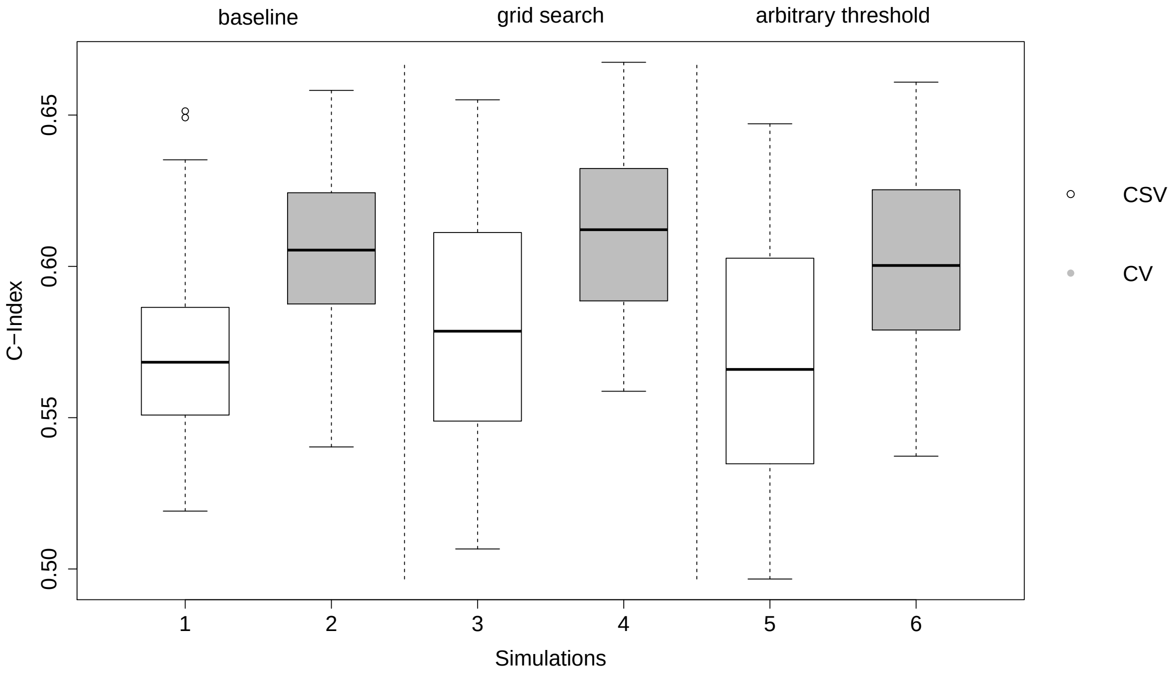


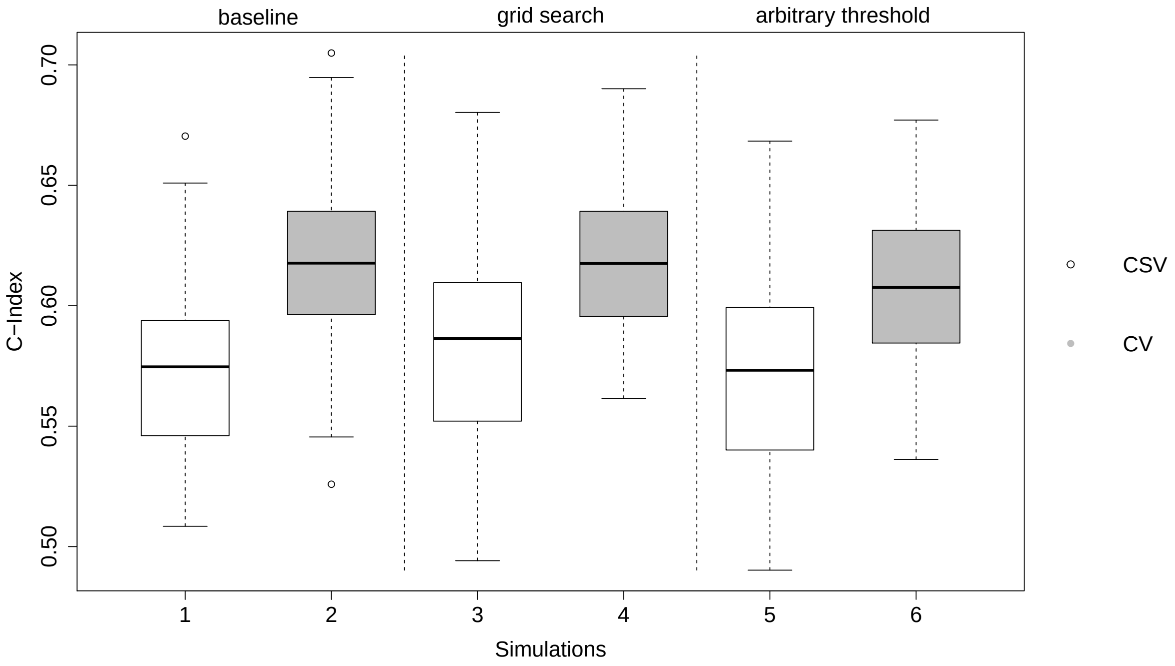
Figure 5: Spaghetti plots of averaged C-indices in cross-study validation and cross-validation. For each set of simulations where a specific source of heterogeneity is equalized, we took the 100 matrices of C index and computed the average values on every position in the matrix. Each line represents the C indices on a certain position (diagonal for CV, off-diagonal for CSV) in the matrix, which changes along with the simulations regarding different sources. The six sets of simulations on the x-axis correspond to the panels in Figure 1. 1: baseline simulation, 2: eliminating heterogeneity in clinical covariates, 3: filtering genes with high Integrative Correlation, 4: enforcing identical true models, 5: Using only identical baseline hazards but different coefficients. 6: altering all factors together. The order of CSV scores is shuffled from panel 1 where no source is changed, to panel 2 where the covariates are balanced, especially for the breast cancer / Más-o-menos combination.

8 Fixed threshold for filtering genes with Integrative Correlation

Breast Cancer / Más-o-menos



Breast Cancer / Ridge Regression



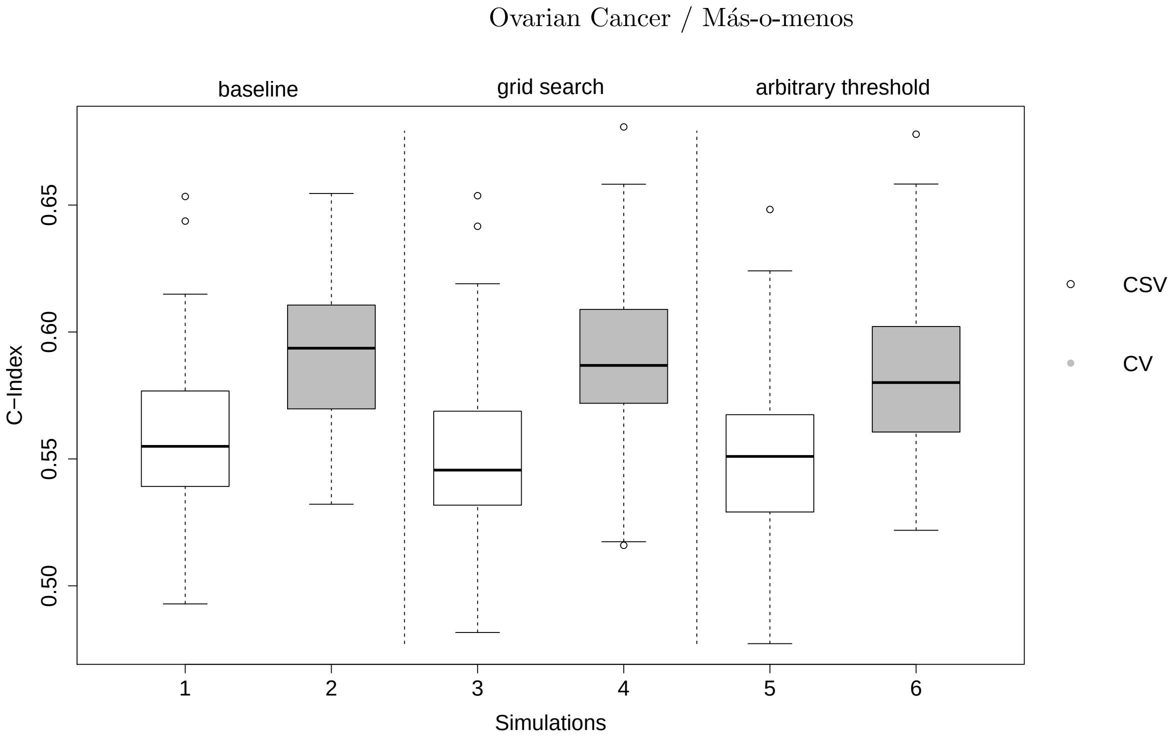
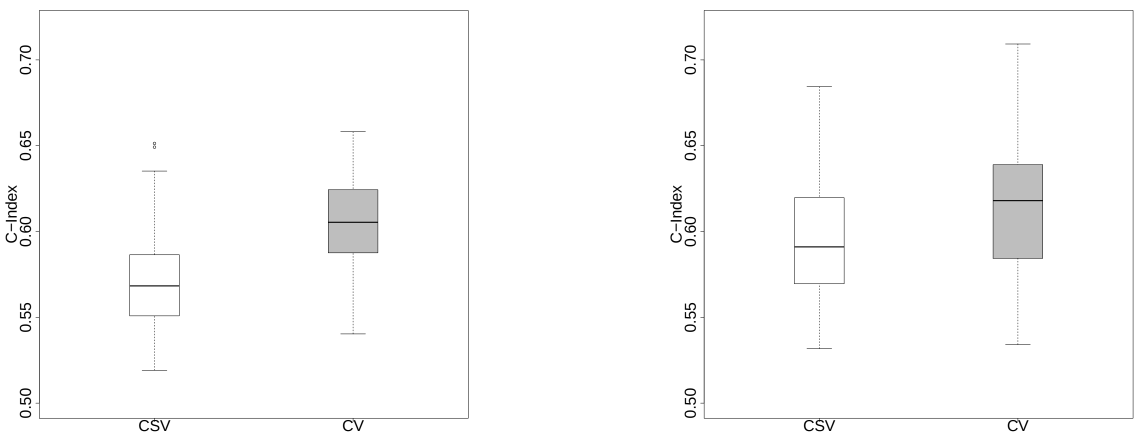


Figure 6: Boxplots comparing the impact of different thresholds of Integrative Correlation on the simulation result. Baseline: simulation where no sources of heterogeneity are altered, which corresponds to panel "baseline" in Figure 1. Grid search: we searched on grid for the threshold such that the top 1000 genes with high Integrative Correlations are selected. Arbitrary threshold: the cutoff values of 0.4 for breast cancer and 0.2 for ovarian cancer are used, which includes 343 out of 5307 genes for breast cancer datasets, and 590 out of 7484 genes for ovarian cancer datasets. Both CV and CSV scores generated by using fixed thresholds tend to be worse than baseline, while using the top 1000 genes tends to give better result. The difference between CV and CSV is not reduced with either approach.

9 Remove CAL



(a) Include CAL

(b) Remove CAL

Figure 7: Comparison before and after removing CAL under the baseline condition. No factors are considered in this plot. The reduction of difference in C-index between CSV and CV is quite noticeable after eliminating CAL, thus motivates us to repeat the whole simulation without CAL to check if there is any difference in the result. This will be shown in Figure 8.

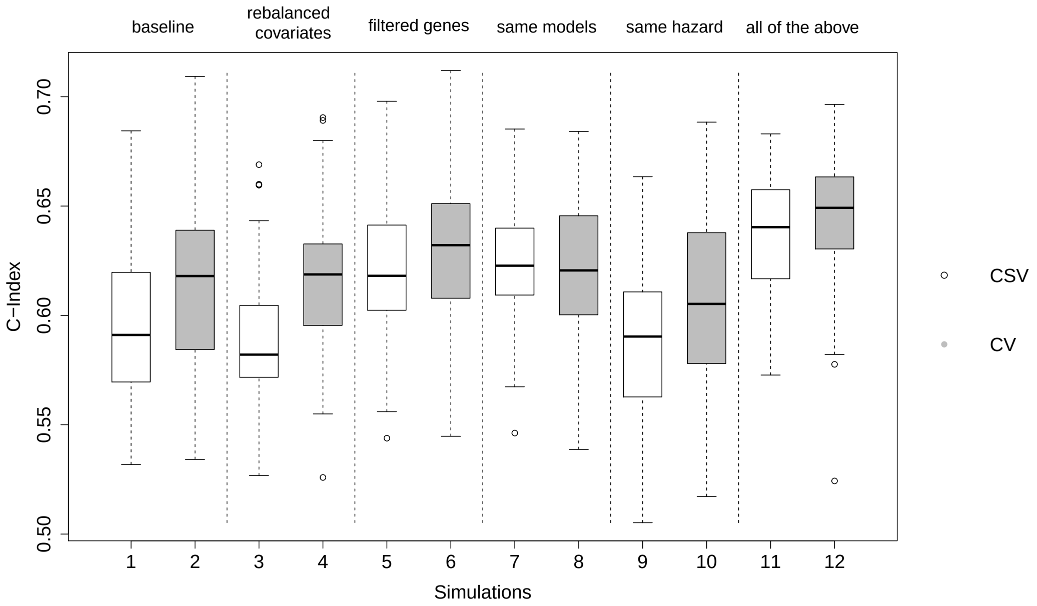


Figure 8: Simulation Results with CAL removed. Removing CAL improves both CV and CSV, and it affects CSV more than CV. The absolute value of the CV-CSV difference is reduced in almost every simulation. However, compared to Figure 1, the results stay robust after the outlier data set is removed, which rules out its influence on the conclusions.