BIOINFORMATICS

Vol. 00 no. 00 2014
Pages 1–9

Dataset heterogeneity in the validation of prediction models across studies

Yuqing Zhang [1,4], Christoph Bernau [2,3], Giovanni Parmigiani [4,5] and Levi Waldron [6] *

[1]Boston University Bioinformatics Program, Boston, U.S.A
[2]Leibniz Supercomputing Center, Garching, Germany
[3]Department for Medical Informatics, Biometry and Epidemiology, Munich, Germany
[4]Dana-Farber Cancer Institute, Boston, U.S.A
[5]Harvard School of Public Health, Boston, U.S.A
[6]School of Urban Public Health at Hunter College, City University of New York, New York, U.S.A

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Motivation: Cross-study validation (CSV) of prediction models is an alternative to traditional cross-validation (CV) in research domains where multiple comparable datasets are available. Although many studies have noted potential sources of heterogeneity in genomic studies, to our knowledge none have systematically investigated their intertwined impacts on prediction accuracy.

Methods: We employ a hybrid parametric/non-parametric bootstrap method to generate realistic simulations based on publicly available breast and ovarian cancer microarray datasets, where two predictive models are validated within and across studies. Three types of heterogeneity between studies are manipulated and studied: 1) imbalances in the prevalence of clinical and pathological covariates, 2) differences in gene covariance that could be caused by batch, platform, or tumor purity effects, and 3) differences in the "true" model that associates gene expression and clinical factors to outcome, including model coefficients and baseline hazard. We assessed model accuracy by Concordance Index (C-Index) within and across studies while altering these factors and the combinations of them.

Results: Lower accuracy is seen between than within studies, and the difference cannot be explained by heterogeneity in the available clinical covariates or by differences in gene covariance. Forcing identical generative models nearly eliminates the within/across study difference for both cancer types. These results suggest that the most easily identifiable sources of study heterogeneity may not be the primary ones determining the ability to translate prediction models across studies.

Availability: The simulation methodology is implemented as the *simulatorZ: Simulator for Collections of Independent Genomic Datasets* Bioconductor package (http://bioconductor.org/packages/release/bioc/html/simulatorZ.html).

*to whom correspondence should be addressed

1 INTRODUCTION

Quantification of heterogeneity between studies and its impact on validation of decision models is important across a wide range of applications. A special issue of Briefings in Bioinformatics (Validation in Bioinformatics and Molecular Medicine, May 2011) emphasized that independent validation of genomic classifiers is rare (Castaldi *et al.*, 2011), and that difficulty with external validation and study heterogeneity is common not only in microarray studies but extends to GWAS studies (König, 2011). External validation is critical in any research domain affected by heterogeneous samples, sample selection bias, or technical batch effects. However it has proven especially difficult for classifiers and subtypes identified from gene expression data. Patient populations can be heterogeneous in their exposures, geography, race/ethnicity, and socioeconomic status, and these differences could manifest as biologically distinct forms of disease that vary systematically between studies. Adding to the potential for study bias, clinical tissue specimens are costly and difficult to collect, and time-consuming to connect with clinical follow-up. What statisticians call "samples of convenience" are the norm in translational genomics research (Simon *et al.*, 2009), even though convenience hardly describes the process of collecting and assaying clinical specimens. Methods for correction of validation accuracy estimation in biased samples have been proposed (Cortes *et al.*, 2008), but only when the unbiased distribution of covariates is known. In the practice of developing genomic prediction models, all potential sources of heterogeneity are likely not even known. Finally, batch effects (Leek *et al.*, 2010) and platform effects impact on reproducibility across studies, but sources of batch variation often are not known. In spite of these challenges, we expect clinically-relevant genomic findings to be reproducible at hospitals around the world, suggesting robustness in the presence of some heterogeneity. The "molecular portraits" of breast cancer, for example, have been broadly reproduced across platforms and centers (Hu *et al.*, 2006). Heterogeneity between clinical studies cannot be avoided, but we propose that the impact even of still-unidentified sources of

heterogeneity can be accounted for and quantified if independent studies are available.

Previous studies have shown that accuracy estimates of genomic prediction models based on independent validation are inferior to cross-validation estimates (Castaldi *et al.*, 2011; Bernau *et al.*, 2014), but did not identify the sources of heterogeneity responsible. A standard approach is to remove as many sources of heterogeneity as possible, such as Waldron *et al.* (2014) and Riester *et al.* (2014), which limited meta-analyses to late-stage, high-grade, serous ovarian cancer. Similarly, recommendations for the replication of genome-wide association studies include studying a "similar" population. However in many cases it is unclear what measures of study similarity are important, and unnecessarily restrictive inclusion criteria have costs in reduced sample size and loss of generality of findings. Thus the question arises of which sources of heterogeneity do in fact impact the accuracy of cross-study prediction, and how these can be determined from the data when independent studies are available.

We compare within and across-study validation of omics-based prediction models using simulations which are generated from two collections of publicly available experimental datasets, for estrogen-receptor positive breast cancer and for late-stage, high-grade ovarian cancer. We investigate the impact of three possible types of heterogeneity on cross-study validation performance: changes in prevalence of known clinical and pathologic factors, changes in gene expression covariance structure for example due to batch or platform effects, and changes in the true models associating gene expression and clinical factors with outcome. These sources of heterogeneity are manipulated and equalized in turn, while comparing within- to across-study validation of risk scores for survival. The methodology of this study can be applied to investigating the effects of study heterogeneity on model validation in any scenario where multiple independent but comparable datasets are available.

2 METHODS

We evaluate the effects of between-study heterogeneity by resampling of studies, and preserving the distribution and covariance of gene expression through resampling of individual patients within studies. We generate linear models associating clinical/pathological variables and gene expression to outcome, as well as baseline survival functions and censoring distributions, based on the original experimental data. We emphasize that standard clinical factors were included as required, unpenalized covariates in the "true" prognostic model so that their associations with outcome would be guaranteed to be preserved across simulations. We include these clinical covariates as binary predictors, along with gene expression, in within and across-study validation.

We review the simulation procedure in the following sections, which involves a 3-step bootstrap method. To implement this simulation approach, we developed the *simulatorZ* package and made this available through Bioconductor. Scripts for reproducing the results of this paper are stored and documented at https://bitbucket.org/zhangyuqing/datasetheterogeneity.

2.1 Datasets

The simulations were performed on two collections of cancer studies, with the first group being the curated breast cancer datasets used in the work of Bernau *et al.* (2014). This compendium of microarray datasets was originally published by Haibe-Kains *et al.* (2012), with censored disease and metastasis-free survival (DMFS) response and estrogen-receptor positive breast cancer individuals. Datasets originating from more than one study were separated into their base sets. 10 base sets involved in the previous study were considered for inclusion: CAL, MAINZ, MSK, EXPO, TRANSBIG, UNT, VDX, MDA4, SUPERTAM_HGU133A and SUPERTAM_HGU133PLUS2 (the latter four were referred as VDX3, MDA5, and TAM by Bernau *et al.* (2014)). MDA4 and EXPO were excluded for the lack of the DMFS information. As illustrated in Haibe-Kains *et al.* (2012), we linearly scaled the gene expressions based on the 2.5% and 97.5% quantiles. Among genes shared across the remaining 8 datasets, 40% with high variances were selected, which led to 5307 genes. Samples from SUPERTAM_HGU133A duplicated in VDX were removed. Missing expression values were imputed by the K-nearest neighbors approach (Troyanskaya *et al.*, 2001).

Available clinical and pathological covariates in these datasets include PGR and HER2 expressing status, histological grade, tumor size, nodal status, and patient age at diagnosis. Age and tumor size were dichotomized at 50 years and 2cm, respectively (Haibe-Kains *et al.*, 2012), and grade was maintained as three levels (low/medium/high) as in the original datasets. PGR, nodal status and HER2 were available in 5, 4, and 2 datasets respectively. Thus only tumor size, grade and patient age were considered as available clinical factors. Patients missing any of these three clinical factors, or with zero follow-up, were excluded. Supplementary Table 1 shows detailed availability of covariate information in each data set.

Finally, datasets with fewer than 40 individuals were removed. The remaining 7 datasets are referred to as the "original datasets" (Table 1), with 1021 patients combined. Synthesized Hazard Ratios (HR) of the tumor size, grade and patient age are 1.96, 1.65 and 1.2 (Supplementary Table 2). This is consistent with commonly used prognostic factors for primary breast cancer such as the Nottingham Prognostic Index (Haybittle *et al.*, 1982), showing that the dichotomized tumor size and grade are prognostic.

The other collection of ovarian cancer microarray datasets was published in Ganzfried *et al.* (2013) and made available through the *curatedOvarianData* Bioconductor package. To select a group of studies for simulation, we filtered datasets with at least 1000 genes, 40 individuals and 15 events. Batch effects were adjusted within each data set when the batch information was available, and the expression values were standardized with z-score scaling. Missing values in the expression matrices were imputed by the K-nearest neighbors approach (Troyanskaya *et al.*, 2001). We then took the overlapping genes across all studies, and removed all genes with a standard deviation below the 60% quantile in every selected dataset, which resulted in 7484 gene features.

Patients with late-stage, high-grade cancer were included. The datasets contain patient age and debulking as available clinical covariates. Samples with missing values for these covariates or the outcome were removed. The patient age was dichotomized at 70 years. Distributions of these covariates are summarized in Table 2. Synthesized Hazard Ratios of these covariates are 1.84 (age) and 1.48 (debulking), as shown in Supplementary Table 3. Forest plots of HR for each covariate in these two collections of data are provided (Supplementary Figure 2). After cleaning, 5 microarray

NO.	Name	#patients	%>2cm	%high-grade	%medium-grade	%low-grade	%>50yrs	reference
1	CAL	68	63.2	57.4	35.3	7.4	61.8	Chin *et al.* (2006)
2	MNZ	162	40.7	9.9	73.5	16.7	76.5	Schmidt *et al.* (2008)
3	TAM1	317	58.0	18.3	58.4	23.3	91.2	Foekens *et al.* (2006)
4	TAM2	128	53.1	31.3	44.5	24.2	93.0	Symmans *et al.* (2010)
5	TRB	132	43.2	26.5	51.5	22.0	29.5	Desmedt *et al.* (2007)
6	UNT	72	38.9	16.7	43.1	40.3	59.7	Sotiriou *et al.* (2006)
7	VDX	142	2.8	71.8	25.4	2.8	57.7	Minn *et al.* (2007)
	overall	1021	44.1	29.6	50.9	19.5	72.3	

Table 1. Covariate distributions in the ER-positive breast cancer microarray base datasets. Percentages are rounded to the nearest tenth. Datasets acronyms: CAL = University of California, San Francisco and the California Pacific Medical Center (United States), MNZ = Mainz hospital (Germany), TAM1 and TAM2 represents SUPERTAM_HGU133A and SUPERTAM_HGU133PLUS2, which are provided by Haibe-Kains *et al.* (2012), TRB = TransBIG consortium dataset (Europe), UNT = the cohort of untreated patients from the Oxford Radcliffe Hospital (United Kingdom), VDX = Veridex (the Netherlands). Column labels: #patients = the number of patients after cleaning. %>2cm = percentage of patients in cleaned datasets with tumor size larger than 2cm. %low, medium and high-grade refer to percentage of patients with high, intermediate and low level of histological grade respectively. %>50yrs = percentage of patients older than 50 years.

NO.	Name	#patients	%>70yrs	%suboptimal	reference
1	E.MTAB.386	124	29.8	22.6	Bentink *et al.* (2012)
2	GSE26712	182	24.7	51.1	Bonome *et al.* (2008)
3	GSE49997	136	14.7	31.6	Pils *et al.* (2012)
4	GSE9891	124	19.4	42.7	Tothill *et al.* (2008)
5	TCGA	369	23.3	27.6	Network *et al.* (2011)
	overall	935	22.7	34.1	

Table 2. Covariate distributions in the overian cancer microarray base datasets. Column labels: #patients = the number of patients after cleaning. %>70yrs = percentage of patients older than 70 years. %suboptimal = percentage of patients in cleaned datasets with suboptimally debulked tumors, as opposed to the optimally debulked ones.

datasets were selected as the "original datasets" for ovarian cancer, with 935 patients in the collection.

2.2 The *simulatorZ* Bioconductor package

Bernau *et al.* (2014) introduced a systematic approach for synthesizing a group of independent microarray datasets with survival outcome for cross-study assessment of prediction algorithms.

We developed the *simulatorZ* Bioconductor package to create a collection of independent genomic datasets with realistic properties and time-to-event outcome generated from a known risk model. *simulatorZ* also implements the Más-o-menos algorithm (Zhao *et al.*, 2014; Donoho and Jin, 2008) and provides basic facilities for cross-validation and cross-study validation of prognostic models. It supports *ExpressionSet* objects to simulate identifier-based data such as from microarray, and *SummarizedExperiment* objects to simulate genomic range-based data such as from DNA and RNA sequencing.

2.2.1 Simulation Approach The simulation procedure is in three steps. The first is a non-parametric bootstrap at dataset level, in which datasets are sampled with replacement from the list of original datasets. This estimates variability due to sampling of studies from a "super-population" of studies (Hartley and Sielken, 1975). The second step is another non-parametric bootstrap resampling at the patient level, where a number of observations are sampled with replacement from each dataset selected in step 1. In the final step, a proportional hazards model is fit to the *original* datasets, then used to simulate time-to-event on the simulated sets (parametric bootstrap):

$$M^j_{true} : \lambda^j(t|x) = \lambda^j_0(t) * \exp(\beta^T_j x) \quad (1)$$

M^j_{true} is the true PH model for the j-th data set, whose hazard function is expressed with $\lambda^j(t|x)$, with x as covariates. $\lambda^j_0(t)$ is the baseline hazard function for this set. β represents the regression coefficients.

The generative model in step 3 combines the truncated inversion method of Bender *et al.* (2005), the Nelson-Aalen estimator (Nelson, 1969, 1972; Aalen, 1978) for cumulative hazard functions, and the *CoxBoost* method of generating best-fit linear risk scores (Binder and Schumacher, 2008). Detailed methods can be found in the references; here we briefly illustrate the workflow. We first use *CoxBoost* to obtain coefficients of linear predictors fitted to the original base datasets, using the genes plus the clinical covariates, such as tumor size, debulking, histological grade and patient age, as predictors. The prognostic covariates were included to be mandatory unpenalized. We also obtained the Nelson-Aalen estimator of baseline cumulative survival and censoring hazard, which, together with the *CoxBoost* coefficients and equation 1, defined the "true" models of disease-free survival for each dataset. Finally, we have:

$$U = S(T) = \exp[-H_0(T) * \exp(\beta^T x)] \sim Uni[0, 1] \quad (2)$$

H_0 is the baseline cumulative hazard for the lifetime random variable T. S denotes the survival function. We sample two

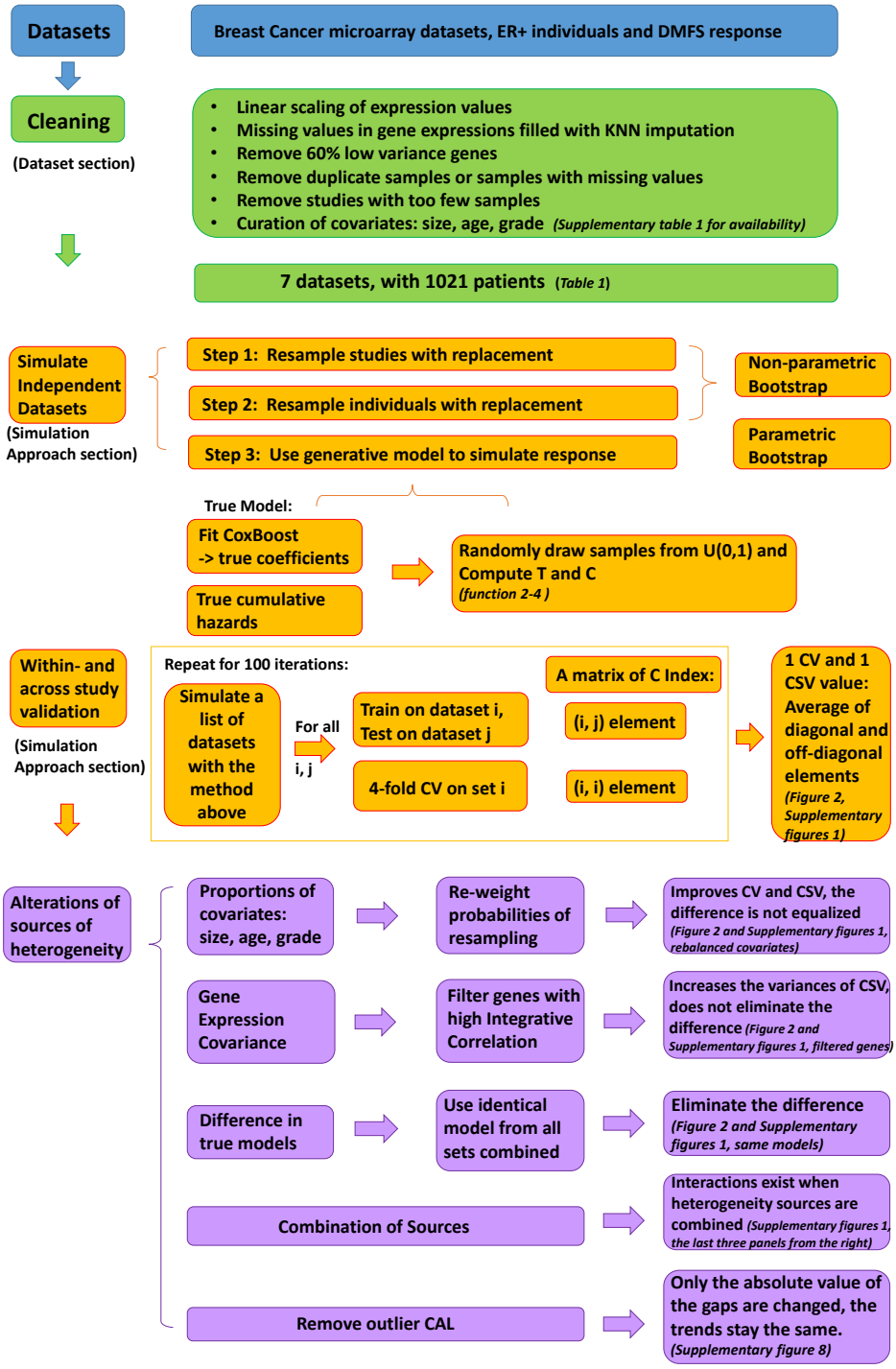

Fig. 1. A schema of our study. Methods and major results for breast cancer are summarized in this flow chart. Equivalent methods were employed for ovarian cancer.

independent, uniformly distributed variables $u1$ and $u2$, then simulate the survival (T) and censoring (C) time, with

$$T = H^{-1}_{surv_0}[-\log(u1) * \exp(-\beta^T x)] \quad (3)$$

$$C = H^{-1}_{cens_0}[-\log(u2) * \exp(-\beta^T x)] \quad (4)$$

$H^{-1}_{surv_0}$ and $H^{-1}_{cens_0}$ are the inverses of baseline cumulative survival and censoring hazard, respectively. These are inverted by finding the point on the time line such that the values calculated

by $-\log(u) * \exp(-\beta^T x)$ are closest in absolute value to the cumulative hazards. The simulated survival response is the smaller one between T and C. This completes the simulation of datasets.

The Más-o-menos algorithm (Zhao *et al.*, 2014; Donoho and Jin, 2008) and ridge regression (Hoerl and Kennard, 1970) were chosen as examples of predictive models to generate risk scores on the simulated datasets. We repeated 100 simulations and model validations for each of the three dataset / algorithm combinations, which are breast cancer data with Más-o-menos and ridge regression methods, and ovarian cancer with the Más-o-menos algorithm. Bernau *et al.* (2014) and Zhao *et al.* (2014) have shown that these algorithms perform comparably to more complicated methods in these datasets.

In each of 100 iterations, we simulated a list of independent datasets of sample size $n = 150$ using the "original sets" and this 3-step bootstrap approach. We then generated a matrix of C-indices for all combinations of training and test sets, as described below (Bernau *et al.*, 2014). Cross study validation (CSV) performance was summarized by the simple average of C-statistics for training and validation across all pairs of independent studies (off-diagonal elements of the matrix), and performance of the 4-fold cross validation (CV) was summarized by the average of the diagonal elements (Bernau *et al.*, 2014). The process was repeated while altering potential sources of across-study heterogeneity, as described in the next three subsections of Methods. The workflow of methods is summarized in Figure 1.

2.3 Clinical Covariates

The proportions of covariates such as tumor size, grade, debulking and of young and old patients were balanced so that on average, each simulated study would have proportions equal to the overall proportions. To balance on multiple covariates, we re-weighted individual bootstrap sampling probabilities of each patient to result in identical joint probability distributions of the clinical / pathologic covariates across datasets. Re-weighting of the sampling probabilities, rather than enforcing strict equality of covariate proportions, reflects the reality that these proportions are subject to sampling variation.

2.3.1 Mixed-Effect Models To quantify the impact of balancing population attributes to the C-statistics in cross-study validation, we built a linear model for each of the clinical factors, treating the changes in covariate distribution in both the training and the test set as fixed effects, with random slopes varying across 100 simulations. Such a linear model is formulated as

$$\Delta C \sim \Delta cov(train) + \Delta cov(test) + interaction \quad (5)$$

For instance, the linear model for patient age at diagnosis is

$$\Delta C \sim \Delta \%old(train) + \Delta \%old(test) + \Delta \%old(train) : \Delta \%old(test) \quad (6)$$

where Δ represents the differences of values computed on simulations with balanced covariates compared to those with the unbalanced ones. C stands for the cross-study validation score calculated from the performance matrices. $\%old(train)$ and $\%old(test)$ are the percentage of old patients in the training and the test set. The last term is the interaction between the train and test terms. For each dataset / algorithm combination, we used the off-diagonal elements in 100 matrices of C-indices to form the design matrix. All terms in the linear model are regarded as fixed effects with random slopes, which avoids treating observations as independent within each simulation.

2.4 Expression Covariance

To investigate the potential impact of heterogeneity between gene expression levels in different datasets, we compared the baseline case to the case where we only use genes with high Integrative Correlation (Parmigiani *et al.*, 2004; Garrett-Mayer *et al.*, 2008) between every dataset pair. Briefly, we first calculated the Pearson correlation matrix of each gene expression matrix. For each pair of datasets, the Pearson correlation of the k-th rows of the two correlation matrices is the Integrative Correlation of gene k. We did a grid search for the threshold of the Integrative Correlation, such that around 1000 genes with the highest Integrative Correlation scores between every pair of *original* datasets were included. We also used arbitrary cut-offs 0.4 for breast cancer and 0.2 for ovarian cancer, as comparison.

2.5 True Models

We equalized the "true models" of each dataset separately in terms of coefficients of the linear risk score for each dataset, and the baseline hazard function. The common true model was fitted using all original datasets combined into a single large dataset. However this model achieved much lower within-study performance than the baseline simulation. To increase CV to the same average C-Index as achieved without equalized coefficients, we multiplied all equalized coefficients by a constant factor of 1.6 in all dataset / algorithm combinations, that resulted in CV performance similar to the baseline simulation.

3 RESULTS

In original and simulated data, we observed dramatic loss of survival discrimination accuracy in cross-study validation (CSV) when compared to cross-validation (CV). These reductions were approximately 0.04 on the C-index scale. We manipulated aspects of the simulated data to establish that reducing heterogeneity in prevalence of important clinical covariates and in gene expression measurements were not sufficient to eliminate the loss of cross-study prediction accuracy relative to within-study accuracy. Enforcing identical "true models" of survival as predicted by gene expression and clinical covariates was sufficient to largely reduce and even eliminate the between to within-study gap in prediction accuracy.

3.1 Simulation Studies

We re-capitulated the major properties of two collections of cancer microarray studies in a 3-step bootstrap simulation procedure.

The simulation resulted in within and across-dataset training/ validation characteristics comparable to those from prior clinical studies (Supplementary Figure 3). Most importantly, the simulation studies maintained a marked difference in prognostic validation accuracy as estimated within studies and across studies by the

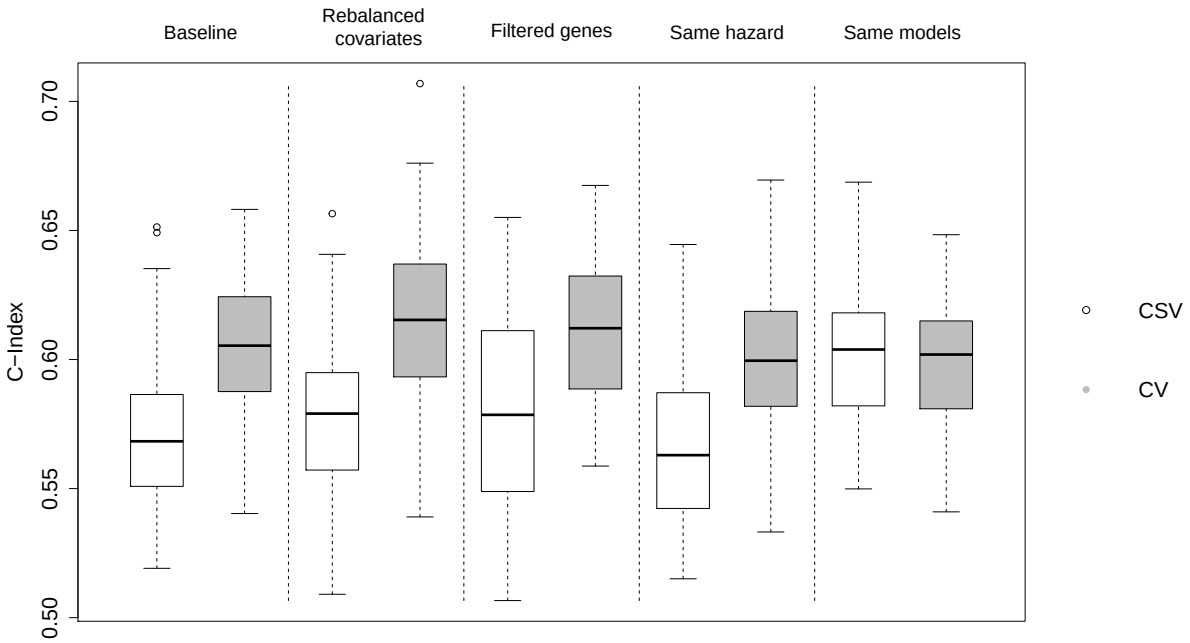

Fig. 2. Simulation results comparing performances of CV and CSV for training Más-o-menos on breast cancer data. Panel "baseline" is a baseline scenario without considering any source of heterogeneity. In "rebalanced covariates", we change the re-sampling probability to coalesce with the distribution of covariates. The differences between CV and CSV are not eliminated. Panel "filtered genes" considers around 1000 genes with high Integrative Correlation, which increases the variance of CSV, but does not explain the difference as well. Panel "same hazard" separates the "true" model, i.e., the same coefficients from *CoxBoost* and the cumulative hazard, and uses the same cumulative hazard but different coefficients for simulation. "Same models" shows great change using both the same coefficients and the same cumulative hazard. A comparison between the last two panels suggests that true coefficients is mainly responsible for the change in "same models". This figure is our major result which shows that the same true model can largely mitigate and even remove the differences between CSV and CV, while the other two factors cannot explain the performance drop of CSV. Results for the other two dataset / algorithm combinations, and for altering multiple sources, are displayed in Supplementary Figure 1.

C-index. This difference is seen for ridge regression and the Más-o-menos method, in breast cancer and ovarian cancer, see panels "baseline" in Figure 2 and Supplementary Figure 1.

3.2 Eliminating Heterogeneity in Clinical Factors

To establish whether eliminating heterogeneity in known clinical factors could improve cross-study validation accuracy, we re-weighted bootstrap sampling probabilities to balance tumor size, grade and patient age for the breast cancer data, and age and debulking status for the ovarian cancer data. Proportions of these factors were then the same, on average, for each simulated dataset. For both cancer types, the differences between CV and CSV are not eliminated (Figure 2 and Supplementary Figure 1, "rebalanced covariates"). In this scenario, differing distributions of these covariates across datasets does not contribute to the loss of prediction accuracy across studies relative to within studies.

Though the re-balancing of covariates does not mitigate the overall drop in accuracy in cross-study validation, we found that the rank of CSV scores may be affected depending on the algorithm used. Supplementary Figure 5 displays re-ordering of performance ranks for different training / test pairs, but with no overall effect across all training / testing pairs.

To quantify this observation, we used a linear model for every covariate which associates the changes in proportions of clinical factors to the changes in the CSV scores. Table 3 summarizes the results of this analysis and highlights which covariates have an effect on CSV in the two methodologies used. Interestingly, even covariates that are prognostic of the survival outcome do not necessarily significantly relate to the prediction accuracy changes.

3.3 Filtering Genes by Integrative Correlation

We searched on grid for the threshold of Integrative Correlation (Parmigiani *et al.*, 2004; Garrett-Mayer *et al.*, 2008) to include roughly the top 1000 genes with the highest Integrative Correlation between every pair of original datasets. The threshold is 0.24 for breast cancer (999 genes), and 0.15 for ovarian cancer (1002 genes). After filtering these genes on the original datasets, the variance of C-statistics across simulations increases, but the differences between CV and CSV are not equalized.

As a comparison to grid search, we used arbitrary thresholds of 0.4 for breast cancer and 0.2 for ovarian cancer. Using only genes with Integrative Correlation greater than 0.4 reduced the number of genes from 5307 to 343 (6.5%) for breast cancer. A threshold of 0.2 reduced the number of genes from 7484 to 590 (7.9%) in ovarian

Breast Cancer			
		Más-o-menos	Ridge Regression
age	train	0.026 *	0.012
		(0.011)	(0.010)
	test	0.023 *	0.027 *
		(0.011)	(0.012)
	interaction	-0.086 .	0.008
		(0.049)	(0.046)
size	train	-0.036 **	-0.003
		(0.013)	(0.019)
	test	-0.024 .	0.022
		(0.014)	(0.023)
	interaction	-0.119	-0.173
		(0.108)	(0.146)
grade	train (high)	-0.039 *	-0.026
		(0.018)	(0.020)
	test (high)	-0.087 ***	5e-5
		(0.022)	(0.026)
	interaction (high)	-0.170 *	0.020
		(0.071)	(0.059)
	train (mid)	-0.013	-0.020
		(0.020)	(0.028)
	test (mid)	-0.070 *	0.003
		(0.029)	(0.034)
	interaction (mid)	0.170	-0.048
		(0.108)	(0.108)
Ovarian Cancer			
		Más-o-menos	
age	train	-0.014	
		(0.023)	
	test	-0.036	
		(0.028)	
	interaction	0.586	
		(0.511)	
debulking	train	0.029 .	
		(0.015)	
	test	-0.003	
		(0.014)	
	interaction	-0.089	
		(0.145)	
Significant Codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1			

Table 3. Regression table of the mixed-effect models. Train, test and interaction refer to the changes in the proportion of the covariates as illustrated in equation 5. Regression coefficients are estimated by treating each term as fixed effect with a random slope. The values in parentheses report the standard errors of the coefficients. Significance of these coefficients are indicated based on the code in the table.

cancer data. Both CV and CSV performs slightly worse using the fixed thresholds compared to the grid search, which results from the loss of good predictors due to strict filtering. The observed pattern remains (Supplementary Figure 6), which suggests that the result is robust to the two choices of thresholds. Filtering genes to enforce similar covariance structures across studies, as would be expected in the absence of microarray batch or platform effects, reduces but does not remove the CV-CSV difference.

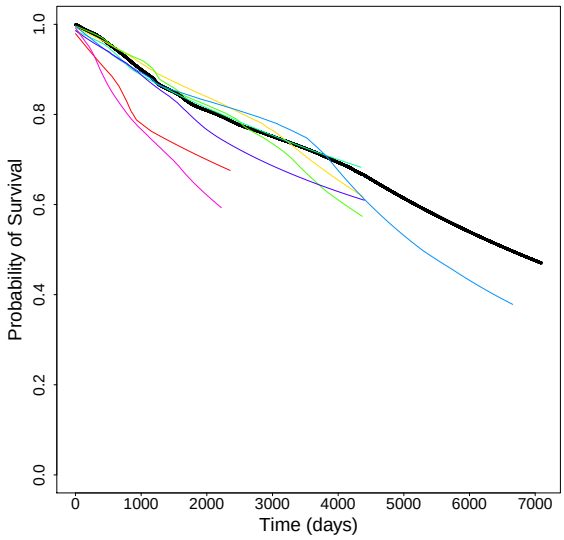

Fig. 3. Average probability of survival of each set and the combination of all sets for breast cancer. The true models are different in true baseline cumulative hazard and coefficients. We can compute a survival function for each individual in every dataset with the true cumulative hazard and the linear predictor. We then average these survival functions across patients within each dataset. Colored lines represent average survival functions in each original sets, while the black, bold line shows the average survival function of all sets combined, with which we will later obtain the same true model in the simulation.

3.4 Using the Same True Model

True models of experimental sets differ in both baseline survival and coefficients. Figure 3 shows the average probability of survival in each dataset and for all datasets combined. Differences in baseline survival mean that survival across all patients is better in some datasets than others, for example varying between 60% and nearly 90% 5-year survival. We equalized both coefficients and baseline hazards across studies, while allowing the joint distribution of covariates and the gene covariance structure to remain heterogeneous.

Utilizing true survival models that are identical in both baseline survival and coefficients is sufficient to eliminate the difference between within and across-study validation for breast cancer (Figure 2 and Supplementary Figure 1, "same models"), whereas equalizing only the baseline survival functions has negligible effect (Figure 2 and Supplementary Figure 1, "same hazard").

Enforcing the identical true models does not completely erase the differences between CV and CSV for the ovarian cancer data, but the differences are greatly reduced compared to baseline simulation.

3.5 Combination of factors

When using identical models, balancing the clinical factors and selecting genes with high Integrative Correlation at the same time, we found that the combined alteration almost removes the CV-CSV differences. A small gap was re-introduced in between the averaged CV and CSV scores compared to using only identical true

models (Supplementary Figure 1, "all of the above"). However this difference is much smaller than baseline.

We investigated the impact of the clinical factors and the gene covariance given that the true models are the same. The combination of re-balancing covariates and forcing identical models is more related to the reappearance of the CSV performance drop in "all of the above" compared to using filtered genes and the same true models (Supplementary Figure 1, "same models covariates" and "same models filtered genes"). These results suggest that, other than the impact of single sources, the interaction of different sources could also contribute to the prediction performance discrepancy within and across studies.

3.6 Eliminating Outlier Datasets

Supplementary Figure 3 identifies the CAL dataset as an outlier with much greater difference between cross-validation and cross-study validation than other breast cancer datasets in simulation, and with C-indices of approximately 0.48 when it is used for training, validation, or cross-validation in original experimental data. To establish the impact of this outlier dataset, we repeated all simulations with it removed (Supplementary Figure 8).

Removal of CAL improves the performances of both CV and CSV. CSV is more affected by this outlier dataset. We observed that CSV performances increase to a greater extent compared to the CV performances in all simulations except for the one where identical true models are used. Thus the prediction accuracy drop is mitigated for the rest of the simulations. However, results under every condition still follow the same trends as illustrated above.

4 DISCUSSION

It is commonly assumed that heterogeneity in experimental platforms or procedures, and differences in patient cohorts, compromise the comparability of independent datasets and the application of omics-based prediction models across studies. This could potentially be addressed by minimizing potential sources of heterogeneity, for example by enforcing precise inclusion criteria for patient inclusion. However, such narrowing has costs in sample size and potential generalizability of findings. To the best of our knowledge, no study has suggested a systematic approach to assessing the impact of suspected sources of heterogeneity on across-study performance of prediction models.

When training and validating microarray-based survival models in two compendia consisting of 7 breast cancer datasets and 5 ovarian cancer datasets, we observed a discrepancy in C-index for models validated in fully independent studies when compared to standard cross-validation. Independent validation statistics were 0.04 worse on the C-index scale for independent validation when compared to cross-validation, a sizeable difference that questions the utility of cross-validation for deciding whether to pursue further development of a prediction model developed and validated on a single dataset. We thus investigated the contributions of known and unknown sources of heterogeneity to this discrepancy.

In simulations mimicking these two compendia of datasets, decreasing heterogeneity in important clinical covariates did not reduce the discrepancy between CV and CSV. This finding highlights that it should not be assumed that known differences in the composition of different cohorts will negatively impact the application of prediction models across them, or that stricter inclusion criteria will improve the models.

Mixed-effect models were used to associate the changes in proportions of clinical covariates to the accuracy changes in cross study validation. Interestingly, we found that although some covariates strongly affect survival, they do not have much impact on the cross-study stability of predictions of survival. Heterogeneity in the prevalence of covariates like debulking and patient age for ovarian cancer can impact overall survival, but not the ability to predict overall survival.

Similarly, in these compendia of datasets spanning at least 11 different labs and 4 different microarray technologies, enforcing good expression measurement comparability through selection of genes with high Integrative Correlation (Parmigiani *et al.*, 2004; Garrett-Mayer *et al.*, 2008) only marginally improved the comparability of cross-study validation and cross-validation. Only ensuring fully identical models of association between gene expression and outcome for each study was sufficient to eliminate this discrepancy. Thus in these datasets, the most important sources of heterogeneity from the perspective of cross-study validation are not recorded in the published datasets.

This study has several limitations. We focus mainly on comparisons between single heterogeneity sources. Altering multiple factors yields results that have a less clear interpretation. However, we report these results in the supplement to show the interaction and dynamics between different factors for those interested to explore. We performed simulations only in the context of ER-positive breast cancer and late-stage, high-grade ovarian cancer, using only two prediction algorithms. But our proposed approach shows how the impact of known sources of study heterogeneity can be assessed for their impact on prediction modeling, and that the most obvious heterogeneity may not be the most important. We could only analyze clinical heterogeneity that was available in sufficient numbers of these datasets: patient age, tumor size and grade for breast cancer; age and debulking status for ovarian cancer. By publishing the *simulatorZ* Bioconductor package, which automates all steps of these simulations including covariate balancing, as well as a code repository to reproduce the results of this paper, we hope to encourage further investigation of the effects of study heterogeneity in other predictive modeling contexts.

5 ACKNOWLEDGEMENT

This work was supported by grants from the National Cancer Institute at the National Institutes of Health (1RC4CA156551-01 and 5R01CA142832 to GP, and 5R03CA191447-02 to LW).

REFERENCES

Aalen, O. (1978). Nonparametric inference for a family of counting processes. *The Annals of Statistics*, **6**(4), 701–726.

Bender, R., Augustin, T., and Blettner, M. (2005). Generating survival times to simulate cox proportional hazards models. *Stat. Med.*, **24**(11), 1713–1723.

Bentink, S., Haibe-Kains, B., Risch, T., Fan, J.-B., Hirsch, M. S., Holton, K., Rubio, R., April, C., Chen, J., Wickham-Garcia, E., *et al.* (2012). Angiogenic mrna and microrna gene expression signature predicts a novel subtype of serous ovarian cancer. *PloS one*, **7**(2), e30269.

Bernau, C., Riester, M., Boulesteix, A.-L., Parmigiani, G., Huttenhower, C., Waldron, L., and Trippa, L. (2014). Cross-study validation for the assessment of prediction algorithms. *Bioinformatics*, **30**(12), i105–i112.

Binder, H. and Schumacher, M. (2008). Allowing for mandatory covariates in boosting estimation of sparse high-dimensional survival models. *BMC Bioinformatics*, **9**, 14. CoxBoost.

Bonome, T., Levine, D. A., Shih, J., Randonovich, M., Pise-Masison, C. A., Bogomolniy, F., Ozbun, L., Brady, J., Barrett, J. C., Boyd, J., *et al.* (2008). A gene signature predicting for survival in suboptimally debulked patients with ovarian cancer. *Cancer research*, **68**(13), 5478–5486.

Castaldi, P. J., Dahabreh, I. J., and Ioannidis, J. P. A. (2011). An empirical assessment of validation practices for molecular classifiers. *Brief. Bioinform.*, **12**(3), 189–202.

Chin, K., DeVries, S., Fridlyand, J., Spellman, P. T., Roydasgupta, R., Kuo, W.-L., Lapuk, A., Neve, R. M., Qian, Z., Ryder, T., *et al.* (2006). Genomic and transcriptional aberrations linked to breast cancer pathophysiologies. *Cancer cell*, **10**(6), 529–541.

Cortes, C., Mohri, M., Riley, M., and Rostamizadeh, A. (2008). Sample selection bias correction theory. *CoRR*, **abs/0805.2775**.

Desmedt, C., Piette, F., Loi, S., Wang, Y., Lallemand, F., Haibe-Kains, B., Viale, G., Delorenzi, M., Zhang, Y., d'Assignies, M. S., *et al.* (2007). Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the transbig multicenter independent validation series. *Clinical cancer research*, **13**(11), 3207–3214.

Donoho, D. and Jin, J. (2008). Higher criticism thresholding: Optimal feature selection when useful features are rare and weak. *Proceedings of the National Academy of Sciences*, **105**(39), 14790–14795.

Foekens, J. A., Atkins, D., Zhang, Y., Sweep, F. C., Harbeck, N., Paradiso, A., Cufer, T., Sieuwerts, A. M., Talantov, D., Span, P. N., *et al.* (2006). Multicenter validation of a gene expression–based prognostic signature in lymph node–negative primary breast cancer. *Journal of clinical oncology*, **24**(11), 1665–1671.

Ganzfried, B. F., Riester, M., Haibe-Kains, B., Risch, T., Tyekucheva, S., Jazic, I., Wang, X. V., Ahmadifar, M., Birrer, M. J., Parmigiani, G., Huttenhower, C., and Waldron, L. (2013). curatedOvarianData: clinically annotated data for the ovarian cancer transcriptome. *Database*, **2013**, bat013.

Garrett-Mayer, E., Parmigiani, G., Zhong, X., Cope, L., and Gabrielson, E. (2008). Cross-study validation and combined analysis of gene expression microarray data. *Biostatistics*, **9**(2), 333–354.

Haibe-Kains, B., Desmedt, C., Loi, S., Culhane, A. C., Bontempi, G., Quackenbush, J., and Sotiriou, C. (2012). A three-gene model to robustly identify breast cancer molecular subtypes. *J. Natl. Cancer Inst.*, **104**(4), 311–325.

Hartley, H. O. and Sielken, Jr., R. L. (1975). A "Super-Population viewpoint" for finite population sampling. *Biometrics*, **31**(2), 411–422.

Haybittle, J., Blamey, R., Elston, C., Johnson, J., Doyle, P., Campbell, F., Nicholson, R., and Griffiths, K. (1982). A prognostic index in primary breast cancer. *British journal of cancer*, **45**(3), 361.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, **12**(1), 55–67.

Hu, Z., Fan, C., Oh, D. S., Marron, J. S., He, X., Qaqish, B. F., Livasy, C., Carey, L. A., Reynolds, E., Dressler, L., Nobel, A., Parker, J., Ewend, M. G., Sawyer, L. R., Wu, J., Liu, Y., Nanda, R., Tretiakova, M., Ruiz Orrico, A., Dreher, D., Palazzo, J. P., Perreard, L., Nelson, E., Mone, M., Hansen, H., Mullins, M., Quackenbush, J. F., Ellis, M. J., Olopade, O. I., Bernard, P. S., and Perou, C. M. (2006). The molecular portraits of breast tumors are conserved across microarray platforms. *BMC Genomics*, **7**, 96.

König, I. R. (2011). Validation in genetic association studies. *Brief. Bioinform.*, **12**(3), 253–258.

Leek, J. T., Scharpf, R. B., Bravo, H. C., Simcha, D., Langmead, B., Evan Johnson, W., Geman, D., Baggerly, K., and Irizarry, R. A. (2010). Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat. Rev. Genet.*, **11**(10), 733–739.

Minn, A. J., Gupta, G. P., Padua, D., Bos, P., Nguyen, D. X., Nuyten, D., Kreike, B., Zhang, Y., Wang, Y., Ishwaran, H., *et al.* (2007). Lung metastasis genes couple breast tumor size and metastatic spread. *Proceedings of the National Academy of Sciences*, **104**(16), 6740–6745.

Nelson, W. (1969). Hazard plotting for incomplete failure data. *Journal of Quality Technology*, **1**, 27–52.

Nelson, W. (1972). Theory and applications of hazard plotting for censored failure data. *Technometrics*, **14**, 945–965.

Network, C. G. A. R. *et al.* (2011). Integrated genomic analyses of ovarian carcinoma. *Nature*, **474**(7353), 609–615.

Parmigiani, G., Garrett-Mayer, E. S., Anbazhagan, R., and Gabrielson, E. (2004). A cross-study comparison of gene expression studies for the molecular classification of lung cancer. *Clin. Cancer Res.*, **10**(9), 2922–2927.

Pils, D., Hager, G., Tong, D., Aust, S., Heinze, G., Kohl, M., Schuster, E., Wolf, A., Sehouli, J., Braicu, I., *et al.* (2012). Validating the impact of a molecular subtype in ovarian cancer on outcomes: a study of the ovcad consortium. *Cancer science*, **103**(7), 1334–1341.

Riester, M., Wei, W., Waldron, L., Culhane, A. C., Trippa, L., Oliva, E., Kim, S.-H., Michor, F., Huttenhower, C., Parmigiani, G., and Birrer, M. J. (2014). Risk prediction for late-stage ovarian cancer by meta-analysis of 1525 patient samples. *J. Natl. Cancer Inst.*, **106**(5).

Schmidt, M., Böhm, D., von Törne, C., Steiner, E., Puhl, A., Pilch, H., Lehr, H.-A., Hengstler, J. G., Kölbl, H., and Gehrmann, M. (2008). The humoral immune system has a key prognostic impact in node-negative breast cancer. *Cancer research*, **68**(13), 5405–5413.

Simon, R. M., Paik, S., and Hayes, D. F. (2009). Use of archived specimens in evaluation of prognostic and predictive biomarkers. *J. Natl. Cancer Inst.*, **101**(21), 1446–1452.

Sotiriou, C., Wirapati, P., Loi, S., Harris, A., Fox, S., Smeds, J., Nordgren, H., Farmer, P., Praz, V., Haibe-Kains, B., *et al.* (2006). Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. *Journal of the National Cancer Institute*, **98**(4), 262–272.

Symmans, W. F., Hatzis, C., Sotiriou, C., Andre, F., Peintinger, F., Regitnig, P., Daxenbichler, G., Desmedt, C., Domont, J., Marth, C., *et al.* (2010). Genomic index of sensitivity to endocrine therapy for breast cancer. *Journal of clinical oncology*, **28**(27), 4111–4119.

Tothill, R. W., Tinker, A. V., George, J., Brown, R., Fox, S. B., Lade, S., Johnson, D. S., Trivett, M. K., Etemadmoghadam, D., Locandro, B., *et al.* (2008). Novel molecular subtypes of serous and endometrioid ovarian cancer linked to clinical outcome. *Clinical Cancer Research*, **14**(16), 5198–5208.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., and Altman, R. B. (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics*, **17**(6), 520–525.

Waldron, L., Haibe-Kains, B., Culhane, A. C., Riester, M., Ding, J., Wang, X. V., Ahmadifar, M., Tyekucheva, S., Bernau, C., Risch, T., Ganzfried, B. F., Huttenhower, C., Birrer, M., and Parmigiani, G. (2014). Comparative meta-analysis of prognostic gene signatures for late-stage ovarian cancer. *J. Natl. Cancer Inst.*, **106**(5).

Zhao, S. D., Parmigiani, G., Huttenhower, C., and Waldron, L. (2014). Más-o-menos: a simple sign averaging method for discrimination in genomic data analysis. *Bioinformatics*.