


Performance evaluation of transcriptomics data normalization for survival risk prediction

Ai Ni  and Li-Xuan Qin

Corresponding author: Li-Xuan Qin, Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, New York, NY 10017. Tel.: 646-888-8251; E-mail: qinl@mskcc.org

Abstract

One pivotal feature of transcriptomics data is the unwanted variations caused by disparate experimental handling, known as handling effects. Various data normalization methods were developed to alleviate the adverse impact of handling effects in the setting of differential expression analysis. However, little research has been done to evaluate their performance in the setting of survival outcome prediction, an important analysis goal for transcriptomics data in biomedical research. Leveraging a unique pair of datasets for the same set of tumor samples—one with handling effects and the other without, we developed a benchmarking tool for conducting such an evaluation in microRNA microarrays. We applied this tool to evaluate the performance of three popular normalization methods—quantile normalization, median normalization and variance stabilizing normalization—in survival prediction using various approaches for model building and designs for sample assignment. We showed that handling effects can have a strong impact on survival prediction and that quantile normalization, a most popular method in current practice, tends to underperform median normalization and variance stabilizing normalization. We demonstrated with a small example the reason for quantile normalization's poor performance in this setting. Our finding highlights the importance of putting normalization evaluation in the context of the downstream analysis setting and the potential of improving the development of survival predictors by applying median normalization. We make available our benchmarking tool for performing such evaluation on additional normalization methods in connection with prediction modeling approaches.

Key words: penalized Cox regression; survival prediction; data normalization; handling effects; microRNA microarray; transcriptomics data

Introduction

Survival analysis plays a foundational role in cancer transcriptomics studies for developing reliable predictors of patient prognosis and treatment response [1–4]. While statistical methods are available to address the issues of high dimensionality and signal sparsity in these studies, research is still lacking on the issue of unwanted data variations associated with disparate experimental handling, which is a pivotal feature of transcriptomics data [5–8]. Many of these studies borrowed existing methods for data normalization that were developed in the setting of differential expression analysis for group comparison, when signals are the difference between group means [9–12]. While the performance of these normalization methods has been

extensively studied in group comparison, little research has been done to reevaluate their performance in the setting of survival analysis, when signals are associated with a censored outcome, partly due to lack of awareness and dearth of benchmarking tools [13–18].

We set out to develop the much-needed benchmarking tool and conduct such an assessment in microRNA microarrays [19, 20]. Our approach leverages a pair of datasets for the same set of tumor samples that we previously collected. One dataset was collected with uniform handling to minimize handling effects; the other was collected without uniform handling and exhibited handling effects [21, 22]. The uniformly handled dataset enabled estimation of the biological effects for each sample, serving as 'virtual samples', and the difference between the two arrays

Ai Ni is an assistant professor in Biostatistics at the Ohio State University. His research interests include the analysis of high-dimensional observational data with survival outcome.

Li-Xuan Qin is an associate member in Biostatistics at Memorial Sloan Kettering Cancer Center. Her current research focuses on the development of benchmark data, analytic methods and computational tools for enabling reproducible statistical translations of cancer genomics data.

Submitted: 8 February 2021; Received (in revised form): 20 May 2021

© The Author(s) 2021. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

for each sample allowed estimation of the handling effects for each array in the nonuniformly handled dataset, serving as ‘virtual arrays’. The virtual samples and the virtual arrays were then reassigned and rehybridized to generate additional data under various scenarios for evaluating normalization methods [21, 23]. In addition, outcome data were simulated by sequentially reallocating the observed outcomes to the virtual samples with probability weighting to achieve a prespecified level of association [24].

In this paper, we report this benchmarking tool based on resampling and our findings on how survival prediction accuracy is impacted by the application of three popular normalization methods: quantile normalization [13], median normalization and variance stabilizing normalization [25]. In addition to post hoc data normalization, we also considered the use of two study designs for assigning arrays to samples: randomized and sorted by survival time (leading to longer survival times being assigned to earlier arrays and shorter survival times to later arrays).

Methods

Collection of the empirical data

Ninety-six high-grade serous ovarian cancer samples and 96 endometroid endometrial cancer samples were collected at Memorial Sloan Kettering Cancer Center between 2000 and 2012. Their microRNA expression levels were measured using Agilent microarrays (Release 16.0, Agilent Technologies, Santa Clara, CA) twice, each with a different experimental design. In the 1st design, the 192 samples were handled by a single experienced technician in one experimental run and assigned to the arrays in a balanced manner via the use of randomization and blocking (each slide of eight arrays serves as an experimental ‘block’). In the 2nd design, the same samples were handled by two technicians in multiple batches over time and assigned to arrays in the order of sample collection. We call the 1st design the ‘uniformly handled’ design and the 2nd the ‘nonuniformly handled’ design. Further details on data collection can be found in the articles by Qin *et al.* [21, 22].

Estimation of biological effects and handling effects from the empirical data

Estimation of biological effects

We used the data collected with the uniformly handled design for the 96 ovarian cancer samples as a best approximate for the biological effects of these samples. We call them ‘virtual samples’.

Estimation of handling effects

Assuming that handling effects were additive, we used the differences between the two arrays for each tumor sample (one from each design) to estimate the handling effects for each array in the nonuniformly handled dataset. We call them ‘virtual arrays’ and split them (by whole slides) into two equal-sized sets, one set with the first 96 arrays for prognosticator training and another with the last 96 arrays for validation. The additivity assumption has been deemed reasonable for microarray data and has been adopted in published methods on microarray data normalization and analysis [26, 27].

Elicitation of regression coefficients

We used the data for the 96 ovarian cancer samples from the uniformly handled design to assess each microRNA’s association with progression-free survival (PFS), an important survival

outcome variable in ovarian cancer [28]. PFS is defined as the time from primary surgery to disease progression, death or loss of follow-up, whichever occurs first. The rate of censoring was 23% (22/96) in our data. Univariate Cox regression analysis showed that six markers have P -values <0.005 and regression coefficients ranging from 0.26 to 0.78 [29]. For the purpose of the simulation study, we generated three sets of ‘true regression coefficients’ from the estimated regression coefficients at three different signal levels to elucidate the impact of true marker effect size on survival prediction.

- (i) Moderate signal: we used the quadruple of the estimated regression coefficients for the six significant markers and zero for the other markers.
- (ii) Weak signal: we randomly chose 30 markers and set their regression coefficients to 0.35 and those of other markers to zero. The L1 norm of the weak-signal coefficient vector equals that of the moderate-signal coefficient vector, so the two vectors have the same ‘total’ effect that was distributed differently among markers—one on a few markers with large effects and the other on many markers with small effects.
- (iii) Null signal: we used a null regression coefficient vector with all zeros as a negative control.

Simulation of survival outcome and array data for prognosticator training and validation

Simulation of survival outcome

We simulated new PFSs for the 96 virtual samples based on the true regression coefficient vectors. To ensure a realistic marginal distribution for PFS without having to arbitrarily assume a parametric distribution, we developed a permutation-based procedure to simulate PFS exhibiting various levels of association with biological effects. This method was inspired by a similar permutation-based method proposed by Heller [24]. It first sorts the observed PFS times in an ascending manner regardless of the censoring status and then sequentially pairs them with the virtual samples starting from the smallest time as follows.

- (i) At PFS time $t_{(j)}$, the probability of choosing virtual sample i from those that have not been chosen is calculated as $\Pr(\text{choose virtual sample } i \text{ at time } t_{(j)}) = p_{ij} = \frac{\exp(x_i \beta_0)}{\sum_{k \in R_j} \exp(x_k \beta_0)}$, where x_i is the biological effect of virtual sample i , β_0 is the true regression coefficient vector, R_j is the set of virtual samples that have not been chosen by $t_{(j)}$.
- (ii) Since $\sum_{i \in R_j} p_{ij} = 1$, the selection of virtual samples at $t_{(j)}$ is determined by a single realization of a multinomial distribution with $n = 1$ and $p_j = (p_{1j}, p_{2j}, \dots, p_{k_j j})$, where k_j is the size of R_j .
- (iii) The above steps are repeated through all sorted PFS times.

We conducted a small simulation to demonstrate that the above procedure leads to the intended association between the covariates and the survival outcome and present the results in the Appendix (Supplementary Figures 4 and 5).

Simulation of array data for prognosticator training

Training microarray data were simulated using a process called ‘virtual rehybridization’, so as to preserve the complex correlation structure of the biological effects and handling effects [21, 23]. Namely, the 96 virtual samples (along with their PFS times) were reassigned to the 96 virtual arrays allocated for

Table 1. Summary of simulation scenarios

Scenario Notation ^a	Handling effects in training data	Handling effects in test data	Handling effects associated with outcome in training data	Handling effects associated with outcome in test data
HE00Asso00	No	No	No	No
HE10Asso00	Yes	No	No	No
HE10Asso10	Yes	No	Yes	No
HE11Asso00	Yes	Yes	No	No
HE11Asso10	Yes	Yes	Yes	No
HE11Asso01	Yes	Yes	No	Yes
HE11Asso11	Yes	Yes	Yes	Yes
HE11Asso1-1	Yes	Yes	Yes	Yes (negatively)

^a‘HE’ stands for handling effects. The 1st and 2nd digits following ‘HE’ indicate presence (‘1’) versus absence (‘0’) of handling effects in training and test data, respectively. ‘Asso’ stands for association with survival outcome. The 1st and 2nd digits following ‘Asso’ indicate the presence (‘1’=positive, ‘-1’=negative) versus absence (‘0’) of association between handling effects and survival outcome in training and test data, respectively.

training, and handling effects for each virtual array were then added to the biological effects of the assigned virtual sample. We considered two scenarios for the reassignment: (1) randomized and (2) sorted by simulated PFS time (leading to longer PFS times being assigned to arrays handled by one technician and shorter PFS times to arrays by the other) so that PFS time was associated with handling effects. To examine whether the magnitude of handling effects impacts the performance of the normalization methods, we also simulated training data with augmented handling effects by tripling the values in the virtual arrays before adding them to the virtual samples.

Simulation of array data for prognosticator validation

Data were simulated either with handling effects (following the same virtual rehybridization method, except that the 96 virtual arrays allocated for validation were used) instead or without handling effects (mimicking a scenario when the validation dataset was of better data quality).

Simulation scenarios

Table 1 summarizes the eight scenarios investigated in our simulations. The scenarios are arranged in such a way that handling effects are increasingly prevalent and involved (i.e. associated with the survival outcome). Each scenario used 400 simulation runs.

Prognosticator training and validation

Array data preprocessing

We assessed three normalization methods: (1) quantile normalization, (2) median normalization and (3) variance stabilizing normalization. The same normalization methods were applied to the training and test data, with the latter in a ‘frozen’ manner [30]. Quantile normalization was applied to the training data using the *normalize.quantiles* function in the R package *preprocessCore*; the quantiles derived from the training data were used to apply frozen quantile normalization to the test data. Similarly, when median normalization was used for the training data, frozen median normalization was applied to the test data. Variance stabilizing normalization was carried out using the *vs2* function in the R package *vs2*. In addition to handling effects adjustment via normalization, the data were additionally pre-processed with log2 transformation and median summarization across replicate probes for each marker [31].

Prognosticator training

We applied two commonly used methods for variable selection to build a multivariate Cox proportional hazard model for PFS prediction: (1) univariate filtering using the per-marker *P*-values and (2) regularized Cox proportional hazard regression. To reduce computational burden and alleviate collinearity in model fitting, we prefiltered the markers using two criteria: (1) high abundance (mean expression on the log2 scale among the 96 samples ≥ 8) and (2) no strong intermarker correlation (Pearson correlation coefficient < 0.9). These criteria were applied to each simulated dataset, and the set of markers that passed the filtering varied across simulation runs. Typically, 100–200 markers remained and entered the model-fitting step.

- (i) In the univariate filtering analysis, we assessed PFS association for each marker using a univariate Cox proportional hazards regression, and selected markers with a *P*-value less than or equal to a given cutoff. The selected markers were then included in a multivariate Cox proportional hazards regression model. The *P*-value cutoff was selected via a grid search from 0 to 0.01 by 0.0005. The value that minimized the Akaike information criterion (AIC) from the multivariate Cox model was selected as the optimal cutoff.
- (ii) In the regularized regression, we first used the univariate Cox regression to select $\lfloor n_0/4 \rfloor$ markers that had the largest partial likelihood, where n_0 is the number of events in the training data and $\lfloor \cdot \rfloor$ denotes the nearest integer, and then performed regularized Cox regression with the selected markers using one of two penalties: (1) the LASSO penalty and (2) the adaptive LASSO penalty [32, 33]. This two-step variable selection strategy has been extensively studied in high-dimensional data literature [34–36]. Six-fold cross-validation was used for selecting the tuning parameters of these penalties.
- (iii) As a reference, we also fitted a nonpenalized multivariate Cox regression model using the true predictive markers (six markers for the moderate-signal model and 30 markers for the weak-signal model), referred to as the oracle method. Although the oracle model is not obtainable in practice, these results are nevertheless revelatory for assessing the impact of handling effects and the performance of data normalization with regard to prediction accuracy.

Prognosticator validation

An ideal approach for validating a prognostication model is to assess its predictive accuracy in an independent test dataset. In

our study, test data shared the same biological effects with training data but differed in handling effects and the random sample-to-PFS pairing. Hence, the test data validation mainly reflected the effectiveness of data normalization and the robustness of the resulted model to handling effects. Harrell's C-index was used to measure prediction accuracy [37]. The mean, 2.5th and 97.5th percentile of the C-index among the 400 runs for each simulation scenario were reported.

To summarize, in this simulation study, we assessed the performance of three methods for data normalization in combination with four approaches for prognostic model building using data generated under eight scenarios of handling-effect pattern and three levels of survival signal strength.

Results

Moderate signal

We present the results of the oracle, LASSO-penalized and univariate-filtering methods in the main text. In addition, we include in the Appendix the result of the adaptive LASSO-penalized method, which is very close to that of LASSO-penalized method (Supplementary Figures 1 and 2).

Oracle method

The simulation result, in terms of prediction accuracy measured by the test data C-index, of the oracle method is presented in Figure 1A. Among the three normalization methods, median normalization was the obvious best performer in all eight simulation scenarios, whereas quantile normalization tended to be the worst (closely following variance stabilizing normalization). In the presence of handling effects, the C-index was between 0.72 and 0.82 for median normalization, between 0.68 and 0.75 for variance stabilizing normalization and between 0.64 and 0.73 for quantile normalization, depending on the specific handling-effect pattern. As a reference, when handling effects were absent in both training and test data, the C-index was 0.92, 0.80 and 0.74 for these three methods, respectively. Across the eight scenarios, as handling effects became more prevalent and outcome associated, their negative impact on prediction became stronger. The most influencing factor was the method of normalization, followed by the presence of handling effects and then by their level of outcome association (comparing HE10Asso00 with HE11Asso00 versus with HE10Asso10).

Penalized regression method

Figure 1B shows the simulation result when the LASSO-penalized Cox regression was used to build the prediction model. The relative performance of the three normalization methods stayed similar to that for the oracle method. That is, median normalization was the best and quantile normalization the worst across the eight patterns of handling effects. Their C-index ranged from 0.74 to 0.83, 0.69 to 0.76 and 0.71 to 0.76 for median, quantile and variance stabilizing normalization, respectively, in the presence of handling effects in training data and/or test data; it was 0.92, 0.81 and 0.84 for the three methods, respectively, in the absence of handling effects. Compared with the oracle method, the LASSO method slightly improved the prediction accuracy by up to 0.07 across the simulation scenarios. This is likely due to the fact that LASSO tends to select more predictor markers into the final model than the true model.

Univariate filtering method

Figure 1C presents the results when the univariate filtering method was used to build the prediction model. The relative performance of the normalization methods was again consistent with the oracle method. In the presence of handling effects, the C-index ranged from 0.74 to 0.81, 0.64 to 0.74 and 0.70 to 0.76 for median, quantile and variance stabilizing normalization, respectively. In the absence of handling effects, it was 0.93, 0.88 and 0.90, for the three methods, respectively. Compared with the LASSO method, the prediction accuracy for the univariate filtering method was slightly worse (by up to 0.05) and substantially more variable.

In addition to the above results, we also performed simulations when the magnitude of handling effects was tripled in the training and test data. As expected, the prediction performance of the normalization methods worsened across all scenarios. Nevertheless, the relative performance of these methods remained the same. We therefore did not include the results in the paper.

Weak signal

The simulation results of the three prediction modeling methods under weak signals are presented in Figure 2. They are very similar to those under moderate signals, supporting the robustness of our findings in terms of the performance of normalization methods to the size of predictive signal for survival outcome.

Null signal

We further examined the performance of data normalization under the null model. Noting that the oracle method is not available under the null model and the simulation results for the other two methods were very similar, we present the result for the LASSO method in Figure 3 and that for the univariate-filtering method in the Appendix (Supplementary Figure 3). Under the null model, as expected, the prediction model did not offer any value beyond a random guess regardless of the choice of normalization method in the first six scenarios. However, in the 7th scenario when handling effects existed in both training and test data with positive outcome-association (HE11Asso11), normalization led to a small improvement in prediction with C-index slightly over 0.5; conversely in the 8th scenario when the direction of association between handling effects and the survival time was opposite (HE11Asso1-1), normalization slightly harmed prediction with C-index below 0.5, suggesting that the prediction model performed even worse than a random guess. The observations in the last two scenarios showcased that handling effects that confound with the survival time can either induce a false positive predictor or dampen a true positive one, depending on the direction of the confounding association.

Discussion

Effective prognostic biomarkers of patient clinical outcomes are of keen interest in cancer research, as they can help identify high-risk population, tailor treatment options for patients and design clinical trials for assessing new therapies. While proven useful for biomarker discovery and patient classification in the setting of group comparison, quantile normalization performed poorly for survival outcome prediction in our study. Intuitively, quantile normalization replaces the ranked data of a marker by the averaged ranked data across samples: for the purpose of biomarker discovery, it is effective for removing bias due to

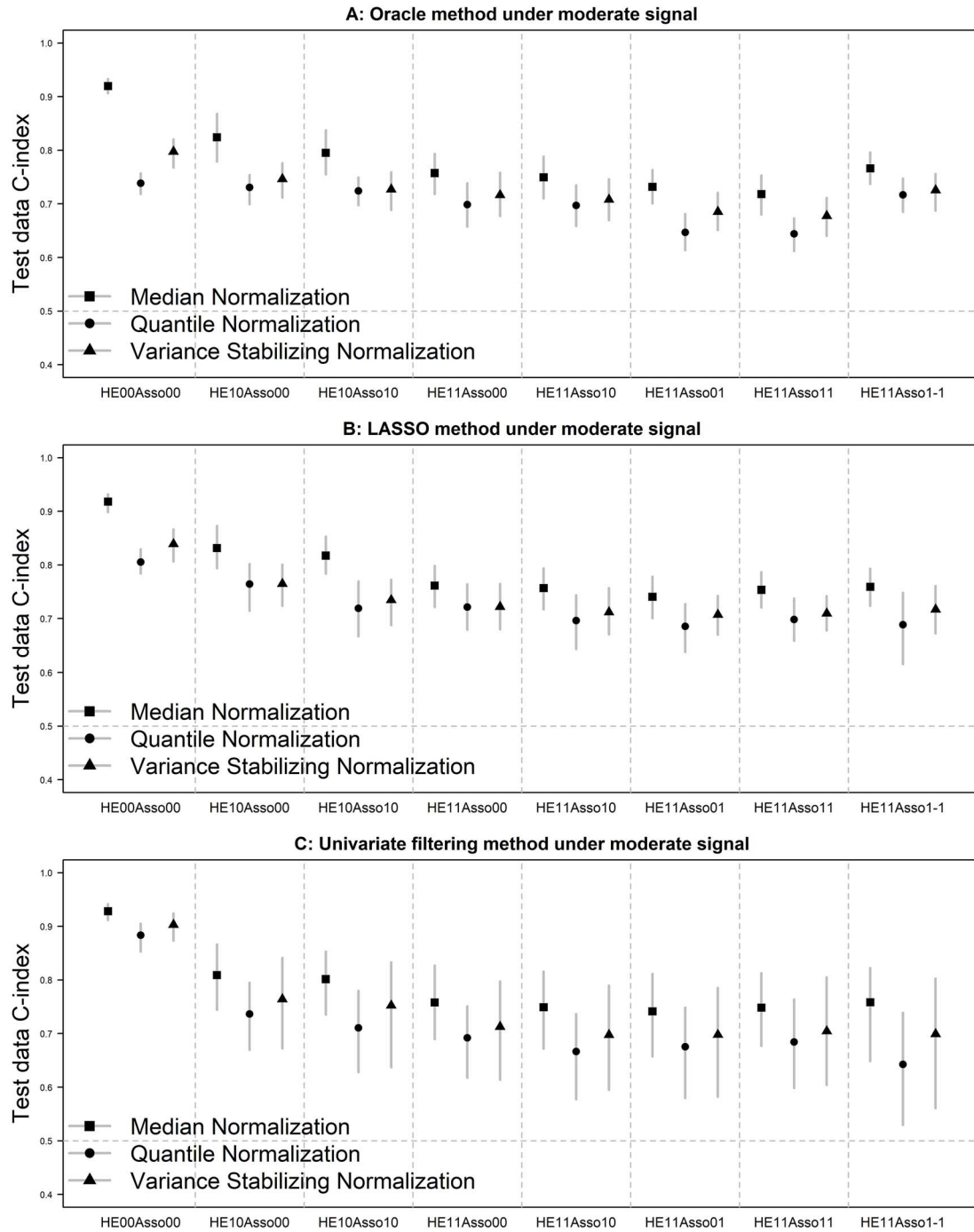


Figure 1. Test data Harrell's C-index of the prediction model developed by the oracle (panel A), LASSO penalized (panel B) and univariate filtering (panel C) method under a moderate level of signal. Vertical bars represent 2.5th and 97.5th percentiles. Symbols in the bars represent mean values.

handling effects in the estimation of group mean difference; for the purpose of building a prediction model for time to event variables, however, it runs the risk of changing the rank of marker data across samples, which we show with numerical examples in the Appendix (Supplementary Figure 4), and subsequently attenuating its regression coefficient toward zero, similar to the effect of adding noise to a predictor [38].

Median normalization, a runner-up in popularity to quantile normalization, performed better than quantile normalization for

the purpose of survival outcome prediction. Its better performance may be explained by the fact that median normalization distorts the rank of marker data to a lesser extent as it only forces the median instead of all percentiles to be the same across samples (Supplementary Figure 5).

The prediction performance of variance stabilizing normalization is only slightly better than quantile normalization. It involves a step where all markers of each sample are rescaled by a sample-specific affine-linear transformation [25]. This

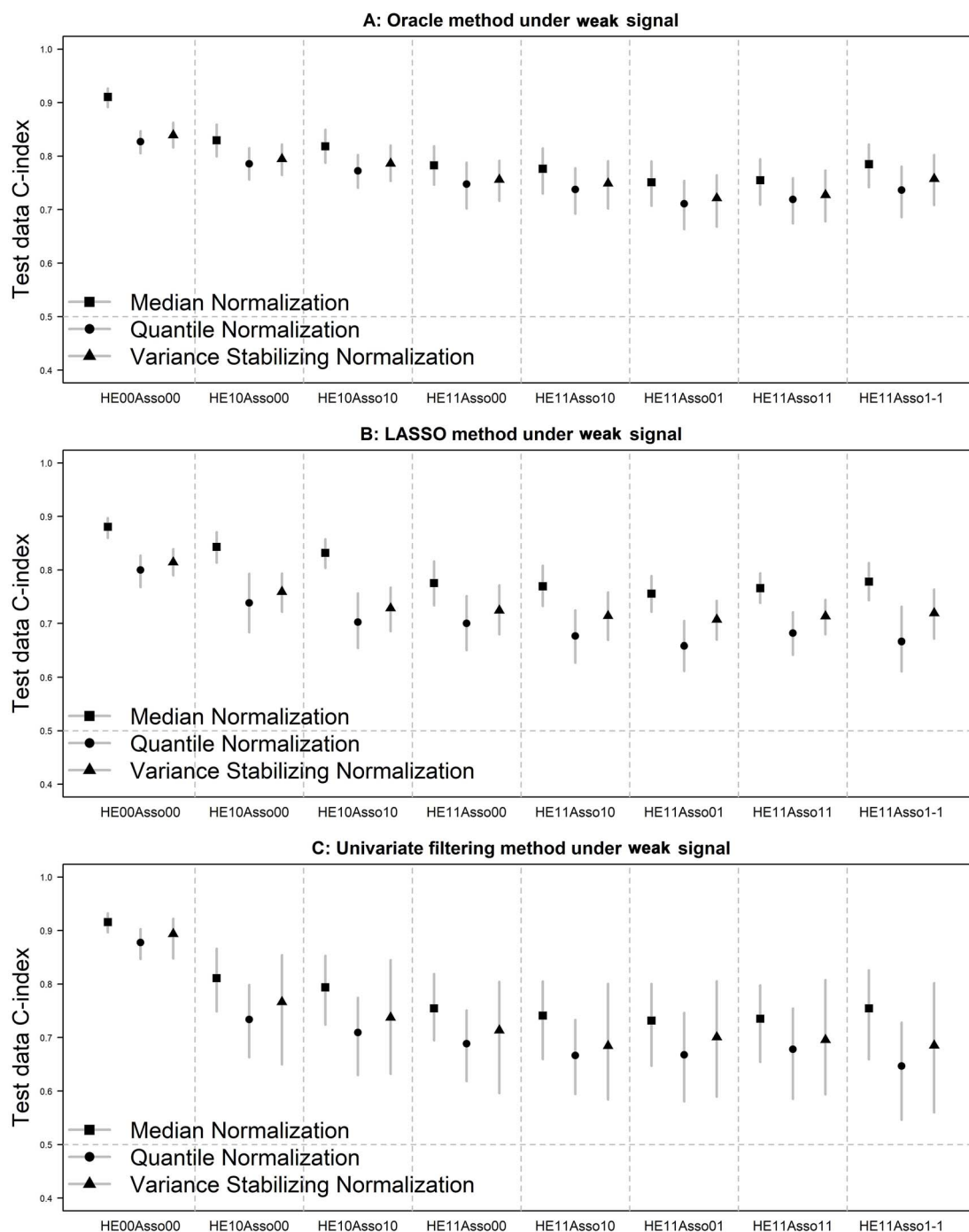


Figure 2. Test data Harrell's C-index of the prediction model developed by the oracle (panel A), LASSO penalized (panel B) and univariate filtering (panel C) method under a weak level of signal. Vertical bars represent 2.5th and 97.5th percentiles. Symbols in the bars represent mean values.

transformation may have resulted in rank distortion of marker data across samples similar to the effect of quantile normalization, hence the unsatisfactory performance in the setting of survival risk prediction.

Our simulation results revealed that penalized regression methods offer slightly more accurate and substantially less variable prediction than the univariate filtering method. This observation agrees with the general opinion in statistical literature that automated significance-based stepwise variable selection

procedures are unstable, especially when the correlation among predictors is high [39]. In our study, adaptive LASSO method tended to give slightly sparser models than the LASSO method, but their prediction performances were similar.

Balanced sample assignment (via the use of study design elements such as blocking, randomization and stratification) has been shown to be effective for avoiding the negative impact of handling effects when developing a predictor of a binary outcome [21]. For predicting time to event outcomes, blocking

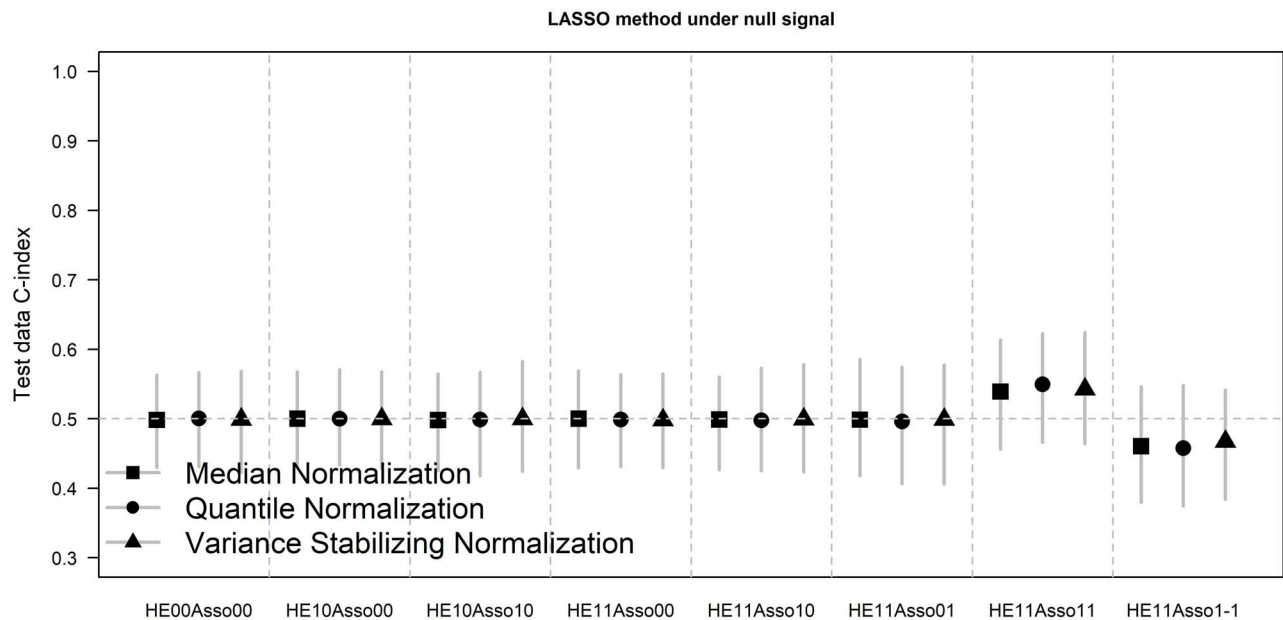


Figure 3. Test data Harrell's C-index of the prediction model developed by the LASSO penalized method under null signal. Vertical bars represent 2.5th and 97.5th percentiles. Symbols in the bars represent mean values.

and stratification are no longer applicable; randomization (i.e. random reassignment of virtual samples to virtual arrays) is still useful as shown in our study. Our study under null signals clearly demonstrated that, in the absence of randomization, spurious positive or negative predictive value (Figure 3 HE11Asso11 and HE11Asso1-1) could arise as a result of the association between survival outcome and handling effects.

In this study we developed the resampling-based simulation strategies and applied them to investigate the performance of data normalization methods for managing handling effects. The same strategies can also be used to study the performance of other categories of methods for managing handling effects, such as the 'Batch-Effect Correction' methods when data are collected in separate 'batches', a notable member of which is ComBat [40]. We will report the findings on the latter alongside a novel alternative method for managing batch effects in a separate paper.

To summarize, our study demonstrates the importance of evaluating the performance of normalization methods in the setting of survival prediction and provides a benchmarking tool for such evaluation for microRNA microarrays. Among the methods examined in this study, median normalization and penalized regression offer better survival risk prediction. We encourage interested researchers to use our tool for assessing additional methods for data normalization and prediction modeling that they use in their practice.

Key Points

- Median normalization offers better survival prediction accuracy than quantile normalization and variance stabilizing normalization.
- In the absence of random assignment of samples to arrays, spurious prediction accuracy could arise due to the association between survival outcome and handling effects.

- It is important to reevaluate normalization methods using the benchmarking tool developed in this paper when handling effect contaminated transcriptomics data are used for survival risk prediction.

Supplementary Data

Supplementary data are available online at <https://academic.oup.com/bib>.

Data Availability

Human tumor tissues used in this study were obtained from participants who provided informed consent, and their use in our study was approved by the Memorial Sloan Kettering Cancer Center Institutional Review Board. The R package containing the data and simulation functions used in this article can be freely downloaded at <https://github.com/LXQi/PRECISION.survival>.

Funding

National Institutes of Health (grants CA214845 and CA008748 to L.X.Q.).

References

1. Lee AH. Prediction of cancer outcome with microarrays. *Lancet* 2005;**365**(9472):1685author reply 1686 PMID: 15894094.
2. Li J, Lenferink AE, Deng Y, et al. Identification of high-quality cancer prognostic markers and metastasis network modules. *Nat Commun* 2010;**1**(1):34 PMID: 20975711.
3. Mobadersany P, Yousefi S, Amgad M, et al. Predicting cancer outcomes from histology and genomics using convolutional

- networks. *Proc Natl Acad Sci U S A* 2018;**115**(13):E2970–9 PMID: 29531073.
4. van't Veer LJ, Bernards R. Enabling personalized cancer medicine through analysis of gene-expression patterns. *Nature* 2008;**452**(7187):564–70 PMID: 18385730.
 5. Consortium M, Shi L, Reid LH, et al. The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nat Biotechnol* 2006;**24**(9):1151–61 PMID: 16964229.
 6. Consortium SM-I. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. *Nat Biotechnol* 2014;**32**(9):903–14 PMID: 25150838.
 7. Mestdagh P, Hartmann N, Baeriswyl L, et al. Evaluation of quantitative miRNA expression platforms in the microRNA quality control (miRQC) study. *Nat Methods* 2014;**11**(8):809–15 PMID: 24973947.
 8. Leek JT, Scharpf RB, Bravo HC, et al. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat Rev Genet* 2010;**11**(10):733–9 PMID: 20838408.
 9. Wang Y, Klijn JG, Zhang Y, et al. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet* 2005;**365**(9460):671–9 PMID: 15721472.
 10. Ueda T, Volinia S, Okumura H, et al. Relation between microRNA expression and progression and prognosis of gastric cancer: a microRNA expression analysis. *Lancet Oncol* 2010;**11**(2):136–46 PMID: 20022810.
 11. Cancer Genome Atlas Research N, Ley TJ, Miller C, et al. Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *N Engl J Med* 2013;**368**(22):2059–74 PMID: 23634996.
 12. Schwalbe EC, Lindsey JC, Nakjang S, et al. Novel molecular subgroups for clinical classification and outcome prediction in childhood medulloblastoma: a cohort study. *Lancet Oncol* 2017;**18**(7):958–71 PMID: 28545823.
 13. Bolstad BM, Irizarry RA, Astrand M, et al. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* 2003;**19**(2):185–93 PMID: 12538238.
 14. Bullard JH, Purdom E, Hansen KD, et al. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics* 2010;**11**(1):94 PMID: 20167110.
 15. Dillies MA, Rau A, Aubert J, et al. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Brief Bioinform* 2013;**14**(6):671–83 PMID: 22988256.
 16. Zhou X, Oshlack A, Robinson MD. miRNA-Seq normalization comparisons need improvement. *RNA* 2013;**19**(6):733–4 PMID: 23616640.
 17. Qin LX, Zhou Q. MicroRNA array normalization: an evaluation using a randomized dataset as the benchmark. *PLoS One* 2014;**9**(6):e98879 PMID: 24905456.
 18. Huang HC, Qin LX. Empirical evaluation of data normalization methods for molecular classification. *PeerJ* 2018;**6**:e4584 PMID: 29666754.
 19. Ambros V. The functions of animal microRNAs. *Nature* 2004;**431**(7006):350–5 PMID: 15372042.
 20. Bartel DP. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* 2004;**116**(2):281–97 PMID: 14744438.
 21. Qin LX, Zhou Q, Bogomolny F, et al. Blocking and randomization to improve molecular biomarker discovery. *Clin Cancer Res* 2014;**20**(13):3371–8 PMID: 24788100.
 22. Qin LX, Huang HC, Villafania L, et al. A pair of datasets for microRNA expression profiling to examine the use of careful study design for assigning arrays to samples. *Sci Data* 2018;**5**(1):180084 PMID: 29762551.
 23. Qin LX, Huang HC, Begg CB. Cautionary note on using cross-validation for molecular classification. *J Clin Oncol* 2016;**34**(32):3931–8 PMID: 27601553.
 24. Heller G. Power calculations for preclinical studies using a K-sample rank test and the Lehmann alternative hypothesis. *Stat Med* 2006;**25**(15):2543–53 PMID: 16025543.
 25. Huber W, von Heydebreck A, Sultmann H, et al. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics* 2002;**18**(Suppl 1):S96–104 PMID: 12169536.
 26. Qin LX, Satagopan JM. Normalization method for transcriptional studies of heterogeneous samples—simultaneous array normalization and identification of equivalent expression. *Stat Appl Genet Mol Biol* 2009;**8**:10 PMID: 19222377.
 27. Kerr MK, Martin M, Churchill GA. Analysis of variance for gene expression microarray data. *J Comput Biol* 2000;**7**(6):819–37 PMID: 11382364.
 28. Barlin JN, Yu C, Hill EK, et al. Nomogram for predicting 5-year disease-specific mortality after primary surgery for epithelial ovarian cancer. *Gynecol Oncol* 2012;**125**(1):25–30 PMID: 22155261.
 29. Qin LX, Levine DA. Study design and data analysis considerations for the discovery of prognostic molecular biomarkers: a case study of progression free survival in advanced serous ovarian cancer. *BMC Med Genomics* 2016;**9**(1):27 PMID: 27282150.
 30. McCall MN, Bolstad BM, Irizarry RA. Frozen robust multi-array analysis (fRMA). *Biostatistics* 2010;**11**(2):242–53 PMID: 20097884.
 31. Qin LX, Huang HC, Zhou Q. Preprocessing steps for Agilent MicroRNA arrays: does the order matter? *Cancer Inform* 2014;**13**(Suppl 4):105–9 PMID: 26380547.
 32. Tibshirani R. The lasso method for variable selection in the cox model. *Stat Med* 1997;**16**(4):385–95 PMID: 9044528.
 33. Zou H. The adaptive lasso and its Oracle properties. *J Am Stat Assoc* 2006;**101**(476):1418–29.
 34. Fan J, Lv J. Sure independence screening for ultrahigh dimensional feature space. *J R Stat Soc Series B Stat Methodology* 2008;**70**(5):849–911.
 35. Fan J, Samworth R, Wu Y. Ultrahigh dimensional feature selection: beyond the linear model. *J Mach Learn Res* 2009;**10**:2013–38 PMID: 21603590.
 36. Tamba CL, Ni YL, Zhang YM. Iterative sure independence screening EM-Bayesian LASSO algorithm for multilocus genome-wide association studies. *PLoS Comput Biol* 2017;**13**(1):e1005357 PMID: 28141824.
 37. Harrell FE, Jr, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med* 1996;**15**(4):361–87 PMID: 8668867.
 38. Carroll RJ, Ruppert D, Stefanski LA, et al. *Measurement Error in Nonlinear Models*, 2nd edn. New York: Chapman and Hall/CRC, 2006.
 39. Harrell F. *Regression Modeling Strategies With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*. Switzerland: Springer International Publishing, 2015.
 40. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* 2007;**8**(1):118–27 PMID: 16632515.