Assessing reproducibility and value in genomic signatures

by

Prasad Patil

A dissertation submitted to The Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy

Baltimore, Maryland

March, 2016

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Consider a genomic signature to be a set of genes whose measured expression is transformed into a prediction of an outcome of interest. Such signatures are the bases of a set of FDA-approved medical tests for predicting the risk of distal recurrence and differential survival in breast cancer patients [84, 83, 63]. The goal of these tests is to provide clinicians with an additional piece of prognostic information that may affect their decision making pertaining to the treatment of a breast cancer patient. As it stands, the tests based on these genomic signatures (MammaPrint, Oncotype DX, Prosigna) are not part of the standard of care for a patient, and there are many issues in the translation of these discoveries from bench to bedside that hinder their reliable use [24]. These issues range from insufficiently thorough validation [86], to technical errors or oversights [5, 46], to outright retraction of results [75].

In addition to the lack of impact in clinical practice, the difficulty of translating these genomic discoveries represents uncertainty about the viability of clinical genomics in general. The vast majority of genetic quantities that are routinely evaluated for a patient were discovered and characterized prior to the

era of high-throughput genomics, e.g. [28, 88, 77, 34, 51]. Although there have been one-off successes [89] and a wide range of candidate and pathway discoveries [85], the costs have been great relative to the payout in terms of widespread clinical use [14]. By examining issues at the point of translation, we can begin to provide a clearer picture of what is possible and realistic to accomplish in the clinic with the discoveries that we have from the high-thorughput era.

Here, we examine two prominent issues in the translation of genomic signatures for risk prediction in breast cancer: reproducibility/replicability and assessment of value added. We address questions of reproducibility and replicability at three levels:

1. on the reproducibility of predictions from a genomic signature-based predictor (Chapter 2). In this chapter, we introduce the notion of "test set bias", which occurs when a genomic signature that has been trained on one dataset is applied to make predictions on another dataset. Due to underlying calculations of distance between gene expression profiles in the process of assigning a risk classification for a particular patient, the test set data require pre-processing and normalization prior to application of the predictive model. We show that this can cause the same patient to receive a different classification depending upon the number and makeup of the patients their profile is normalized with, even though there is no underlying biological change in the patient themselves.

2. on the reproducibility of the process of building a predictive model with gene expression data (Chapter 3). We propose one alternative to gene signature building that avoids the issue of test set bias introduced in

Chapter 2. This process uses rank-based features called Top-Scoring Pairs (TSPs) [79], and we describe novel feature selection and model-building approaches to produce fast, interpretable TSP-based gene signatures. We address the issue of reproducibility of the model-building process by describing the `tdsm` R package, which intends to make the general application of a statistical analysis transparent and well-documented. This is accomplished with prefabricated analysis templates written in R Markdown which restrict user manipulation of parameters and settings.

3. on the distinction between reproducibility and replicability as it pertains to scientific studies (Chapter 5). In this chapter, we examine the issue of replicability. While reproducibility refers to applying the same procedure to the same data and producing the same result [67, 68], replicability (of a study) refers to running a new experiment to address the same scientific question and seeing a result consistent with a previous study [4, 42]. We promote the use of 95% prediction intervals as one way to determine if an effect estimated from a study replication is consistent with the effect estimated in the original study. Although this is done with respect to studies in psychology, the conceptual framework regarding what to expect when a study is replicated can be applied to the realm of validation and confirmation studies that are commonly required for genomic signatures.

The second major topic is the question of the value that a prediction from a genomic signature provides. A risk prediction or classification from a genomic test may be inconclusive or may provide information redundant with what a doctor has already surmised from standard clincial quantities. It is well-known

that doctors send out for genomic tests for a variety of reasons and for a large variety of patients [24]. As a result, determining whether the prediction from a test helped the patient is not as simple as looking at the patient's outcome. For example, if the doctor ignored the prediction from the genomic test and the patient had a good outcome, the outcome ought not be attributed to the test result. We therefore address the question of value by considering how much additional information the prediction from a genomic test could provide a clinician conditional on the clinical quantities already at hand. We approach this in the realm of randomized clinical trials (RCTs), where we have more control over experimental design and a direct method for assessing value added. In Chapter 4, we describe a set of RCT simulations based on a real breast cancer dataset. In this setting, we have available a between-arm treatment effect estimator that can yield improved precision by adjusting for predictive covariates at baseline [21]. Through simulation, we determine how much additional precision we gain by adjusting for different sets of baseline covariates as compared to the basic, unadjusted treatment effect estimator. We are able to then approximate how much additional precision we would stand to gain were we to adjust for the prediction from a genomic test in addition to a set of baseline covariates that a clinician would consider. We find that there is minimal additional gain, representing a direct and realistic assessment of the value of a genomic prediction in the RCT setting.

1.1 Acknowledgements

I would like to acknowledge and thank Jeff Leek, Michael Rosenblum, Ben Haibe-Kains, Antonio Wolff, Don Geman, Elana Fertig, Luigi Marchionni, Bahman Afsari, and Liz Colantuoni for their substantial contributions to and help with this work. I list below the chapters that have been submitted or published as manuscripts and any acknowledgements made within each manuscript.

1.1.1 Chapter 2

This work was completed with Pierre-Olivier Bachant-Winner, Benjamin Haibe-Kains, and Jeffrey T. Leek, and is published with the following citation: Patil, P., Bachant-Winner, P. O., Haibe-Kains, B., and Leek, J. T. (2015). Test set bias affects reproducibility of gene signatures. Bioinformatics, btv157.

This study used data generated by METABRIC; we thank the British Columbia Cancer Agency Branch for sharing these invaluable data with the scientific community.

1.1.2 Chapter 4

This work was completed with Elizabeth Colantuoni, Jeffrey T. Leek, and Michael Rosenblum, and has been revised and resubmitted to Contemporary Clinical Trials Communications. An early pre-print appears with the following citation: Patil, P., Rosenblum, M. A., and Leek, J. T. (2015). Measuring the Contribution of Genomic Predictors to Improving Estimator Precision in Randomized trials. bioRxiv, 018168. (note: author order has changed to Patil, P., Colantuoni, E., Leek, J. T., and Rosenblum M. A. in the final submission).

The authors declare that they have no competing interests.

This research was supported by National Institutes of Health grant R01GM105705. This publication's contents are solely the responsibility of the authors and do not necessarily represent the official views of the above agency.

1.1.3 Chapter 5

This work was completed with Roger D. Peng and Jeffrey T. Leek, and has been revised and resubmitted to Perspectives on Psychological Science. An early preprint appears with the following citation: Leek, J. T., Patil, P., and Peng, R. D. (2015). A glass half full interpretation of the replicability of psychological science. arXiv preprint arXiv:1509.08968. (note: author order has changed to Patil, P., Peng, R.D., and Leek, J. T. in the final submission)

Chapter 2

Test Set Bias Affects Reproducibility of Gene Signatures

2.1 Abstract

Motivation: Prior to applying genomic predictors to clinical samples, the genomic data must be properly normalized to ensure that the test set data are comparable to the data upon which the predictor was trained. The most effective normalization methods depend on data from multiple patients. From a biomedical perspective, this implies that predictions for a single patient may change depending on which other patient samples they are normalized with. This test set bias will occur when any cross-sample normalization is used before clinical prediction.

Results: We demonstrate that results from existing gene signatures which rely on normalizing test data may be irreproducible when the patient population changes composition or size using a set of curated, publicly-available breast cancer microarray experiments. As an alternative, we examine the use of gene signatures that rely on ranks from the data and show why signatures using

rank-based features can avoid test set bias while maintaining highly accurate classification, even across platforms.

Availability: The code, data, and instructions necessary to reproduce our entire analysis is available at https://github.com/prpatil/testsetbias.

2.2 Introduction

One of the most common barriers to the development and translation of genomic signatures is cross-sample variation in technology, normalization, and laboratories [53]. Technology, batch, and sampling artifacts have been responsible for the failure of genomic signatures [69, 5], irreproducibility of genomic results [59], and retraction of papers reporting genomic signatures [75]. Even highly successful signatures such as Mammaprint [84] have required platform-specific retraining before they could be translated to clinical use [33]. An under-appreciated source of bias in genomic signatures is test set bias [52]. Test set bias occurs when the predictions for any single patient depend on the data for other patients in the test set. For example, suppose that the gene expression data for a single patient is normalized by subtracting the mean expression and dividing by the standard deviation of the expression across all patients in the test set. Then the normalized value for any specific gene for that patient depends on the values for all the patients they are normalized with. The result is that a patient may get two different predictions using the same data and the same prediction algorithm, depending on the other patients used to normalize the test set data (**Figure 2.1**).

There are many scenarios under which a patient's classification ought to

change: if new information updates or alters the prediction algorithm, or if the raw, biological patient data itself changes. The case we would like to explore is when the gene signature and prediction algorithm are "locked down" and when there is no biological variation in the patient data. We are concerned with how much data transformation due to pre-processing and normalization affects classification. It is our assertion that steps taken to transform patient data for the purposes of *applying* a prediction algorithm should not alter the patient's eventual classification.

Some normalization methods [57, 70, 11] and some batch correction methods [47, 62] have addressed this issue by normalizing each sample against a fixed, or "frozen", set of representative samples. Unfortunately, these approaches can be applied only to specific platforms where large numbers of representative samples have been collected. This is especially relevant when custom chips are designed, as is the case in many clinical applications. There remain a large range of platforms for measuring gene expression in use by researchers [9], and single sample normalization methods are not currently available for many of these platforms. Additionally, methods such as quantile normalization and other forms of data scaling and transformation have become well-known in the field and are often applied as standard steps in a data processing pipeline.

Even if single sample normalization methods were universally available, public measures of gene expression are frequently pre-processed using a range of methods for cleaning, normalization, and analysis, resulting in a range of expression values for the same gene across different platforms [1]. A more tractable solution is to build gene signatures that do not rely on raw gene expression values. We propose using the ranks of genes instead of their raw expression

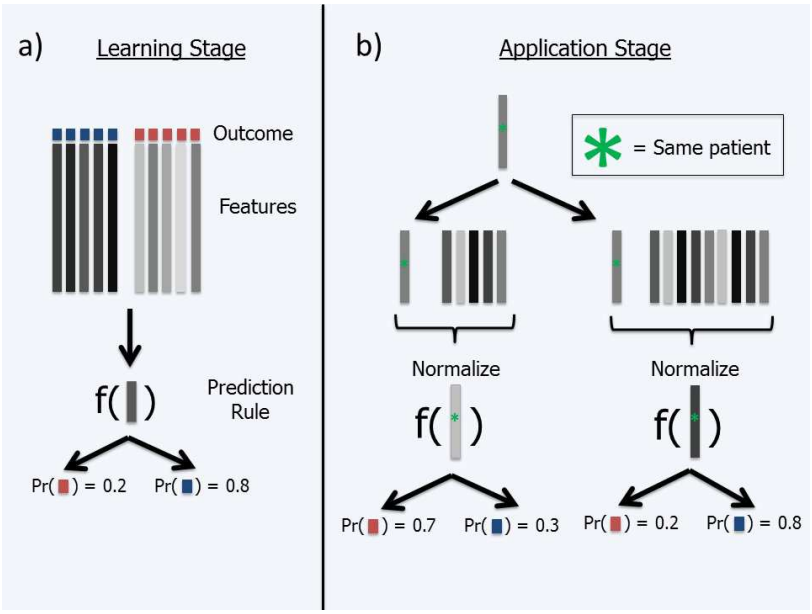

Figure 2.1: A description of how test set bias can alter class prediction for an individual patient. In panel a), we learn a model for predicting if a patient is in class R (red) or class B (blue). In our training data, the patients with darker grey features tend to be in class B, while the lighter grey patients are in class R. We develop a prediction rule from our training data and apply it to a new darker grey patient, and we see that he is likely to be classified to class B. In panel b), we attempt to classify a single patient in the context of two different patient populations. We see that depending on the number and type of other patients in the population when we normalize the data, the resulting feature profile for our patient can be drastically different. This leads to different eventual classifications by our prediction rule. We contend that the ultimate classification of a patient should not depend on the characteristics of the test set, but rather solely on the characteristics of the patient himself.

values under the assumption that any transformation applied to the data is rank-preserving.

As a concrete example, we focus on the PAM50 signature for breast cancer subtyping [63] which is used to assign patients with breast cancer to one of five molecular subtypes: Basal, Luminal A, Luminal B, Her2, Normal. We show that when the number of patients in the test set changes, the predictions for a single patient may change dramatically. We also show that variation in patient populations being predicted upon leads to test set bias. Interestingly, PAM50 can be easily modified into a rank-based signature. We show that predictions from rank-based PAM50 are comparable to those from standard PAM50, and that predictions from rank-based PAM50 are invariant to test set bias.

Test set bias is a failure of reproducibility of a genomic signature. In other words, the same patient, with the same data and classification algorithm, may be assigned to different clinical groups. A similar failing resulted in the cancellation of clinical trials that used an irreproducible genomic signature to make chemotherapy decisions [50]. The implications of a patient's classification changing due to test set bias may be important clinically, financially, and legally. In the example of PAM50, a patient's classification could affect a treatment or therapy decision. In other cases, an estimation of the patient's probability of survival may be too optimistic or pessimistic. The fundamental issue is that the patient's predicted quantity should be fully determined by the patient's genomic information, and the bias we will explore here is induced completely due to technical steps.

2.3 Methods

Study population and data

We collected and curated gene expression microarray data representing 28 independent studies [37]. These datasets spanned 15 different proprietary platform types and a variety of platform versions and included a range of commercial and private manufacturers, spanning Affymetrix, Illumina, and Agilent as well as custom arrays. The data were collected from the Gene Expression Omnibus (GEO) [9], ArrayExpress [64], The University of North Carolina at Chapel Hill database (UNCDB), Stanford Microarray Database (SMD), and Journal and Authors' websites. Metadata were manually curated as previously described [37]. Experiments ranged from 43 to 1,992 patients, with a median of 131 patients and a total of 6,297 patients. (**Table 2.1**).

PAM model fitting

Prediction Analysis of Microarrays (PAM) [82] is a commonly used supervised learning approach for building prediction models using gene expression data from microarrays. We employed the pamr package [40] to fit a PAM model using R. Briefly, pamr takes class labels and microarray data and calculates an average gene expression profile, or centroid, for each class. It then shrinks the centroid to eliminate genes that do not contribute to explaining variability between classes. We then cross-validate to find an appropriate shrinkage threshold to maximize predictive accuracy of our model. We use this threshold to determine how many of the genes to keep in the predictor.

Characteristic	Summary
n	6,297
Age (years)	57.29 (13.42)
RFS (years)	7.22 (4.86)
Tumor Size (cm)	2.52 (1.43)
Node	
+	1,871
−	2,857
NA	1,569
Grade[1]	
1	525
2	1,642
3	2,226
NA	1,904
ER	
+	3,635
−	1,556
NA	1,106
PGR	
+	766
−	656
NA	4,875
Her2	
+	496
−	1,437
NA	4,364
Subtype[2]	
Basal	1,254
Her2	927
LumA	2,007
LumB	1,813
Normal	296

Table 2.1: Baseline characteristics of curated dataset Abbreviations: ER - estrogen receptor status; Her2 - human epidermal growth factor receptor 2 status; Node - whether or not cancer has spread to lymph nodes; PGR - progesterone receptor status; RFS - recurrence-free survival time. Age, RFS, Tumor Size are given as means with standard deviations. [1]due to the ambiguity of grade 2, we chose to build all prediction models for grades 1 and 3 only. [2]subtypes as predicted by PAM50.

Normalization procedure

Normalization is accomplished through quantile rescaling as implemented in the genefu package [39]. This scales each gene expression value x using specific quantiles from the expression data. First, a quantile q is chosen. Through examination of many microarray datasets, $q = 0.05$ was found to be robust. The expression values corresponding to the desired quantiles $q_1 = x_{\frac{q}{2}}$ and $q_2 = x_{1-\frac{q}{2}}$ are defined, and the scaled value $x' = \dfrac{x - q_1}{q_2 - q_1}$ is calculated. In contrast to scaling by the maximum and minimum value, this approach is more robust to extreme outlying gene expression values.

This normalization procedure is applied internally when the intrinsic.cluster.predict function from the genefu package is used and the model's standardization ("std") parameter is set to "robust". For example, we can make PAM50 predictions using pre-packaged models in genefu called pam50 or pam50.robust. The gene centroid information is the same in both cases, but pam50 has std = "none" and pam50.robust has std = "robust". This means that if we apply intrinsic.cluster.predict with pam50, the test data will not be normalized in any way, but if we use pam50.robust the quantile rescaling procedure described above will be applied.

Estimating test set bias

We used two approaches to estimate test set bias. When considering the PAM50 predictor, we simply applied the pre-defined prediction model from the genefu package ([39]) to make predictions on our data.

To train a PAM model, we used 10-fold cross-validation. We create a test

set that is approximately 10% of the total data and use the remaining 90% to train the model. We use the internal cross-validation functions provided in the pamr package [40] to produce a shrinkage threshold and determine the number of genes necessary to make predictions. We then apply this predictor both in the test set, which comes from the same platform, and on other microarray datasets that used different platforms. This process is repeated within each of the cross-validation folds to get average prediction accuracies and standard deviations. When predicting tumor grade (1-3 with increasing severity), we restricted to patients graded 1 or 3 as grade 2 is considered to be ambiguous.

2.4 Results

Normalization makes patient predictions depend on other patients' data

Consider the PAM50 signature [63]. The class assignment for a new patient is made by calculating a measure of closeness between the new patient and the average patient profile in each possible class, then choosing the class that was closest to the sample. For example, PAM50 consists of 50 genes and predicts five classes, so each class centroid is a profile of the average expression of each of the 50 genes within that class. The authors used correlation as a measure of closeness for a given sample to each class centroid - that is, correlation is calculated between the 50 genes in the patient sample and the 50 genes in each class centroid. This is the step that necessitates suitable rescaling of the test data before predictions are made.

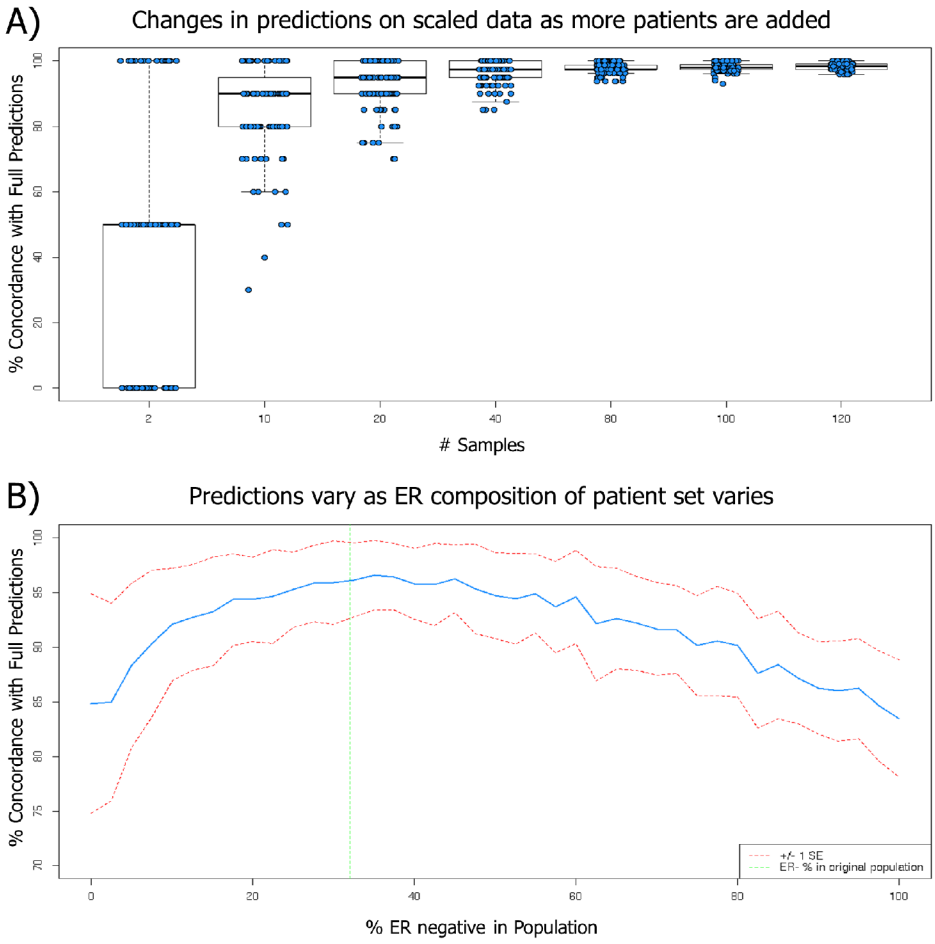

Figure 2.2: Predictions for an individual patient can change depending on how many and what type of patients are included in the normalization step. (A) We first predicted the PAM50 subtype on an entire set of patients (Affymetrix hgu133plus2; GSE7390; n=198). We then took 100 random samples of patient subsets ranging from 2-120 patients and predicted their subtypes with data normalization. We compared this newly predicted subtype to each patient's originally predicted subtype and calculated agreement. Actual data are jittered and overlayed on the boxplot. We find that there is significant variation in percent concordance when a small subset of patients is subtyped in comparison to the entire patient population. (B) From the same setup, we took 100 random samples each of 40 patients and varied the percentage of ER-positive and ER-negative patients in the sample. That is, 0% on the X-axis corresponds to 0% (0/40) ER-negative patients and 100% (40/40) ER-positive patients in the sample. We then predicted subtypes on this subset and compared these newly predicted subtypes to the original predictions. The average concordance is plotted with +/- 1SE bands. We note that the original population is 32% ER-negative (dashed green line), which is where we see close to maximal concordance.

We considered two scenarios which illustrate how PAM50 can produce varying subtype predictions for a particular patient when the data for other patients used in normalization varies. We used data from GSE7390 (n=198), an experiment conducted using the Affymetrix hgu133plus2 microarray. In each experiment, we normalized the gene expression measurements in the test set to fall between 0 and 1.

First we created predictions where we normalized all patients together. Then we calculated predictions for the same patients when normalized in smaller groups (n=2,10,20,40,80,100,120) and measured the agreement between the predictions for the exact same patient when normalized with all patients versus a smaller subset of patients. When normalized in small batches, the predictions for the same patient changed compared to the case where all patients were normalized together **Figure 2.2A**.

Next we predicted on patient populations that varied in the distribution of ER (Estrogen Receptor) status, which is an important factor in breast cancer prognosis and treatment. Again we applied the PAM50 predictor to the entire test set. Then we created subsets of the entire test set with differing percentages of ER-negative patients and applied the predictor to each subset. When the percentage of ER-negative patients in the subset matched the percentage in the entire test set, patient subtypes best agreed with the original predictions on the entire test set. However, when the ER status of the other patients in the test set varied, the predictions for the same patient were often different **Figure 2.2B**.

		Prediction Data											
		Affy				Agilent				Illumina			
	Grade	1		3		1		3		1		3	
Train	Norm.	Mean	s.d.	Mean	s.d.	Mean	s.d.	Mean	s.d.	Mean	s.d.	Mean	s.d.
Affy	Scaled	0.92	0.13	0.67	0.17	0.72	0.05	0.63	0.02	0.95	0.01	0.57	0.02
	Unscaled	0.92	0.13	0.65	0.16	0.79	0.02	0.59	0.01	0.97	0.01	0.51	0.03
Agilent	Scaled	0.93	0.02	0.59	0.04	0.72	0.32	0.56	0.05	0.96	0.01	0.41	0.03
	Unscaled	0.94	0.02	0.55	0.04	0.72	0.32	0.65	0.09	0.97	0.01	0.34	0.05
Illumina	Scaled	0.87	0.00	0.75	0.04	0.75	0.02	0.64	0.01	0.92	0.06	0.65	0.05
	Unscaled	0.79	0.03	0.87	0.02	0.83	0.02	0.58	0.01	0.84	0.08	0.71	0.06

Table 2.2: Average accuracy of scaled and unscaled predictions over different training and testing sets We trained a PAM model to predict tumor grade (either grade 1 or 3) using 10-fold cross-validation on one Affymetrix (GSE7390), Agilent (ISDB10845), and Illumina (ISDB10278) dataset each. The left column presents upon which platform each model was trained, and the top row presents upon which platform each trained model was applied to make predictions. To get average accuracy and standard deviations for a particular platform, we use the model generated under each fold of the cross-validation to make predictions on the remaining test set of the same platform as well as the two other platforms. We applied this model after normalizing ("scaled") the data and after leaving it unnormalized ("unscaled"). We find that the accuracies for predicting grade were similar whether the data were normalized or unnormalized.

Using gene ranks with unnormalized data produces comparable accuracy

When PAM50 was proposed, the authors chose to calculate similarity based on Spearman correlation ([63]). Spearman correlation finds the correlation between the *ranks* of the two sets of gene expression measurements rather than correlation between the actual values. We hypothesized that this rank-based prediction would be immune to some changes of scale across platforms and other platform-specific artifacts. With traditional signatures, these are precisely the reasons why normalization is necessary. To examine this preliminarily, we re-ran the process from the previoius section but simply did not normalize the data and relied on the internal rank-based correlation calculation. We recreated **Figures 2.2A & B** when the data were "unscaled". These appear as **Appendix Figure 7.1**, and they show that the predictions remain constant as sample size and ER status vary when the data are unnormalized and a rank-based metric is employed.

To further evaluate this hypothesis we used the previously proposed PAM signature-building procedure [82] to build a genomic signature to predict tumor grade (a clinical quantity indicating severity) using three datasets measured on different platforms: GSE7390 (Affymetrix; n=198), ISDB10845 (Agilent; n=337), and ISDB10278 (Illumina; n=1,992). We used 10-fold cross-validation to train a model on a particular dataset, made predictions on the testing portion of that dataset, and applied the trained model to the two remaining datasets which represent two different platforms. We averaged over the ten folds in each case to obtain mean accuracy and standard deviation.

To make predictions, we used Spearman correlation to mimic how the PAM50 signature is used [63]. We predicted new patient samples using our PAM signature for grade both with and without normalization. The same set of genes and prediction algorithm are used in both cases - the only difference is that in the former we normalize the test set patient data, and in the latter we leave it unnormalized. We observed that the normalized and un-normalized predictors performed similarly across platforms **Table 2.2**.

Within-platform (Affy-Affy, Agilent-Agilent, Illumina-Illumina in **Table 2.2**), there is no appreciable difference in the average accuracy of predictions when the test data are normalized or unnormalized. For Affy, the grade 1 and 3 average accuracies when the data are normalized are 0.92 (0.13) and 0.67 (0.17), respectively, as compared to 0.92 (0.13) and 0.65 (0.16) when the data are unnormalized. For Agilent, the relevant figures are 0.72 (0.32); 0.56 (0.05) for normalized vs. 0.72 (0.32); 0.65 (0.09) for unnormalized, and for Illumina 0.92 (0.06); 0.65 (0.05) vs. 0.84 (0.08); 0.71 (0.06). In all cases, the ranges of the unnormalized average accuracies substantially overlap those of the normalized average accuracies. Results across platforms (the off-diagonal table entries) tell a similar story. It is the case that if the scaled predictor performs better on grade 1 than the unscaled, then the opposite will be true for grade 3 (see, for example, the Affy-Agilent result). This is due to the fact that patients can be classified as either grade 1 or 3, so if the unscaled version predicts more grade 3 than grade 1, the change in the respective accuracies will be proportional. This analysis suggests that using the PAM predictor for grade with Spearman correlation for making classifications without normalizing the test set data produces similar predictive accuracy to when the test data are normalized.

2.5 Discussion

We found that breast cancer tumor subtype predictions varied for the same patient when the data for that patient were processed using differing numbers of patient sets and patient sets had varying distributions of key characteristics (ER status). This is undesirable behavior for a prediction algorithm, as the same patient should always be assigned the same prediction assuming their genomic data do not change. The fact that sample size affects normalized data values is unsurprising, but the fact that classifications varied by how many patients were ER- in the test set speaks to the generalizibility of an algorithm. Ideally, the test set should be "similar" in composition to the dataset upon which a classification algorithm was trained. The result in **Figure 2.2B** is undoubtedly related to the fact that ER+ patients are different in terms of gene expression from ER- patients, but we see that even slight perturbations in the ER composition of the subpopulation can affect patient classifications. This raises the question of how similar the test set needs to be to the training data for classifications to be trusted when the test data are normalized.

The PAM50 signature uses Spearman correlation to assess distances when making predictions, so we leveraged this by comparing how a PAM signature using Spearman correlation predicts tumor grade outcomes with and without normalization. We found the results to be comparable, but the unnormalized approach guarantees the same prediction for the same patient every time. A gene signature that employs rank-based features or makes other rank-based calculations is one robust approach to avoiding test set bias. Although all gene signature classifiers do not necessarily have a completely rank-based mode as

PAM50 does, the broader implication of this result is that one may endeavour to build predictors that operate on the ranks of data only, thereby bypassing the need for any normalization step when predicting on a test set.

Chapter 3

A Standardized Approach to Building Gene Signatures with Rank-Based Features

3.1 Abstract

The development of a gene signature often involves complicated and irreproducible data modeling and prediction schemes. Here, we introduce a novel gene signature building algorithm and the Templated Deterministic Statistical Machines (`tdsm`) R package. We first describe the statistical underpinnings of an approach that relies on rank-based genomic features with the intention of creating small, interpretable predictive models. We then demonstrate this approach and compose an R Markdown template of the data analysis. We use this template to motivate the use of the `tdsm` package, which promotes transparency and documentation of a statistical analysis by limiting user-driven adjustments.

3.2 Introduction

Apart from a handful of success stories [84, 83, 63] translation of gene signatures from research to clinic has been slower than desired. One primary cause is difficutly with the interpretation and reliability of the underlying predictive model that maps gene expression measurements to an outcome prediction [24]. Another major issue is the lack of reproducibilty plaguing many facets of the signature-building process. We have previously shown that normalization and data pre-processing steps may lead to undesired biases in predictions made by gene signatures [65], which represents an issue with the reproducibility of predictions from a particular gene signature model. Others have described issues with the process of model-building, which propegate in insufficient validation [86], technical oversight [5, 46], and can lead to retraction of seemingly promising genomic predictors [75].

To address reproducibility of predictions, we describe a modeling approach predicated on simple decision trees [81] which use rank-based Top-Scoring Pairs (TSPs) [32] as predictive features. These features are not prone to the normalization and pre-processing issues that may be encountered when dealing with raw gene expression values. We also summarize a novel feature selection scheme that produces relevant and informative gene pairs under very few parametric assumptions.

These methods are bundled into a new R package, `tdsm` (https://github.com/prpatil/tdsm), which uses R Markdown to produce a standard HTML report that describes precisely how the resulting decision tree model is built. Having this report generated every time a gene signature is built will allay some of the

questions surrounding the reproducibility of the model-building process. Written description of each step and the required code are made available within the document, which can be easily shared and examined. We use this templated analysis to build an alternative to the MammaPrint signature which uses fewer genes and a more interpretable model to make predictions of risk of recurrence in breast cancer.

Finally, we describe the structure of the `tdsm` package as it pertains to supporting multiple such templated analyses. We restrict user input to the analysis of choice to only the required data. As a result, users have no control over parameters, which are set by default as part of the analysis template. The user may wish to duplicate a given template and edit parameter choices before running the analysis. For this contingency, we provide `duplicate_template` and `diff_template`, which allow the user to document a comparison between the edited template and the original. The user is therefore informed of how their changes to the template propagated into changes in their results.

3.3 Methods

Suppose we have a training dataset consisting of n patients and m genes whose expression has been characterized. We also have some outcome vector $Y \in \{0, 1\}$ of length n. Similarly, we have a validation dataset consisting of p patients and m genes, with outcome vector $W \in \{0, 1\}$ of length p. We wish to extract some subset $g \in m$ and define a mapping $f : g \mapsto \{0, 1\}$.

3.3.1 Top-Scoring Pairs

The base feature used for prediction is the Top-Scoring Pair (TSP), first suggested by Geman et. al. [32]. For individual i and for two gene expression values, $g_{ij}, g_{ik} \in m$, we may consider the indicator of the expression of the first gene being less than the expression of the second, $z_{ijk} = I(g_{ij} < g_{ik})$. In our setting, the TSP would be the pair of genes that maximizes $|P(g_{ij} < g_{ik}|y_i = 0) - P(g_{ij} < g_{ik}|y_i = 1)|$. In a regression setting, fitting the regression model $E[z_{ijk}|y_i] = \beta_0 + \beta_1 y_i$ and choosing $max_{jk}|\beta_1|$ yields the TSP.

Ideally, we would be able to examine all pairwise comparisons of m genes to choose the TSP, but aside from the computational difficulty, we would find that most gene pairs would not make good predictive features; i.e., most pairs are likely to find $g_{ij} < g_{ik}$ holds for all y_i, hence class differentiation is not possible. Instead, we propose some feature selection approaches that allow us to consider useful pairs in a tractable manner.

3.3.2 Feature Selection - Empirical Controls

To find G we employ a two-step feature selection algorithm. The first step, empirical control feature selection, is a filtering step that does not use the outcome vector Y to pare down the list of candidate gene pairs. As described above, if we consider all pairwise comparisons of m genes most are likely to be unsuitable for differentiating classes, i.e. the vector $Z_{jk} = I(g_{ij} < g_{ik})\ \forall i$ will have zeros or ones in large proportion. Instead, under empirical controls, we search for a specific type of gene pair that has a better chance of "flipping" between the two classes. The pair we desire is one where g_j has fairly constant gene expression

across all individuals (the empirical control), while g_k has high variance in expression across all individuals. **Figure 3.1** displays an example pair of genes that exhibit this relationship.

The procedure to find these candidate pairs is as follows:

1. Sort m genes by their average expression and separate into groups by Q quantiles.

2. Within each quantile grouping, compute and sort by variance and identify the h highest- and lowest-varying genes ($2h$ total genes).

3. Create all possible pairs between the h highest- and h lowest-varying genes.

Figure 3.2 displays the advantage of choosing pairs via the empirical control method. We computed the proportion of ones in the vector Z_{jk} for 6400 randomly selected gene pairs, and did the same for 6400 gene pairs selected via empirical controls. The histogram for the randomly selected pairs has large atoms at zero and one, suggesting that most pairs of genes consist of one gene whose relative expression dominates the other. The histogram for the pairs chosen via empirical controls has a more normal shape. These pairs would be better candidates to associate with an outcome as there is a chance they will differentiate a class due to more consistent "flipping" behavior.

3.3.3 Feature Selection - Conditional Pair Choice

Once we have a candidate set of genes C derived from the empirical control feature selection procedure, we wish to identify any pairs that might be predictive of the outcome. We would additionally like to select subsequent pairs that

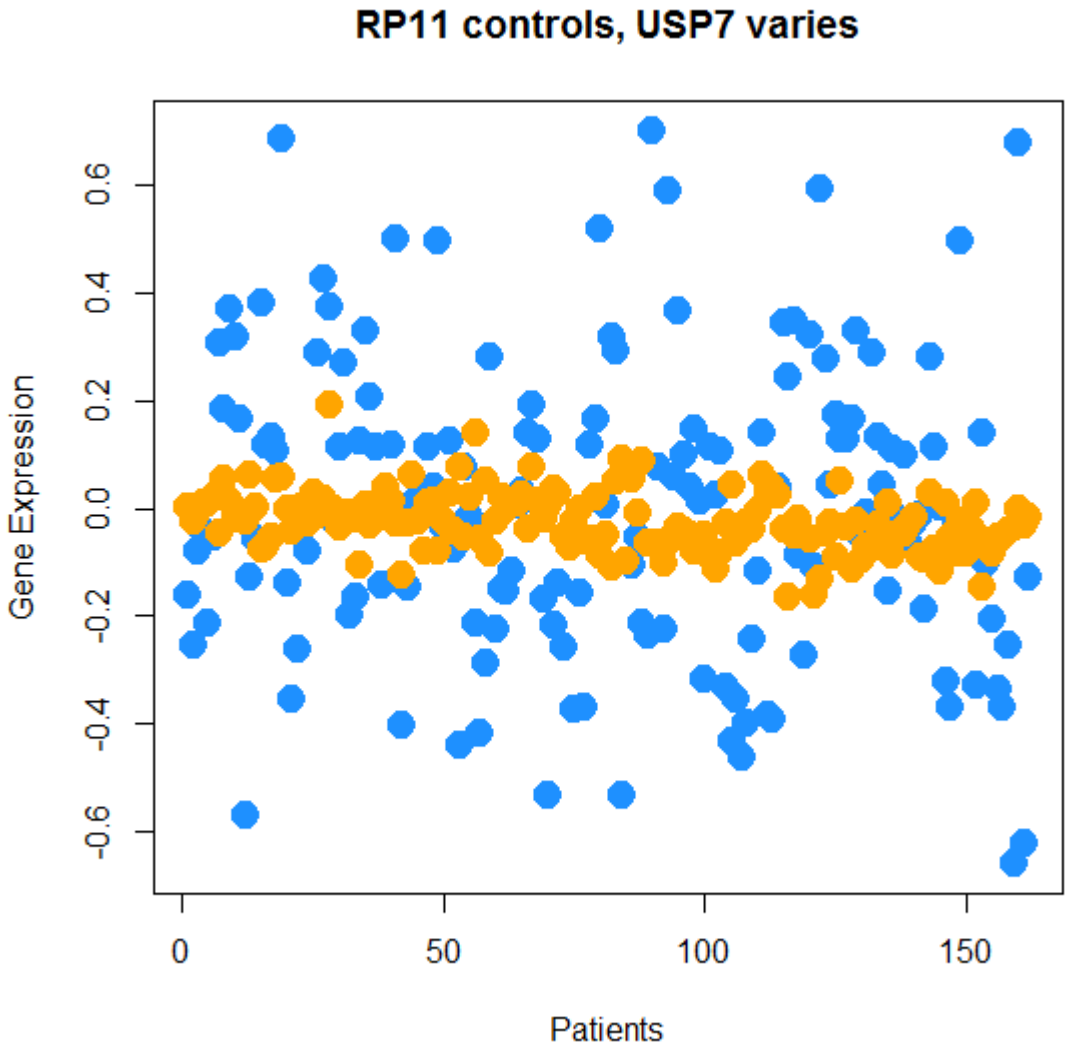

Figure 3.1: Pair of empirical control genes exhibiting "flipping" potential. The raw gene expression values (y-axis) for two genes are plotted for each patient (x-axis). In this case, RP11 (in orange) is the empirical control, low-variancel gene, while USP7 (in blue) is a high-variance gene from within the same quantile as RP11. This pair possesses greater potential for differentiating classes since the relationship between the expression of the two genes is not constant across all patients.

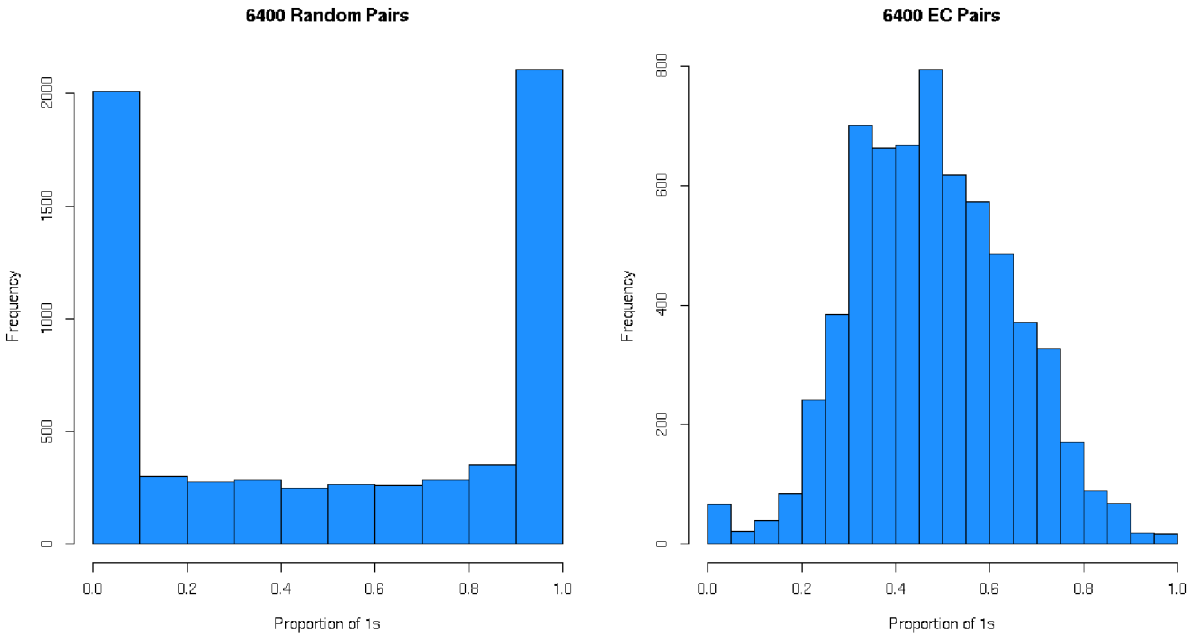

Figure 3.2: 6400 randomly-chosen pairs and 6400 empirical control pairs. On the left is a histogram of the proportion of ones in the vector Z_{jk}, representing the comparison of the expression of genes j and k across all patients, for 6400 randomly chosen genes in a microarray dataset. Note the large percentage of vectors with nearly all zeros or all ones, suggesting that in most cases the expression of one gene tends to dominate the expression of the other across all patients. In contrast, on the right appears a histogram of the same proportion of ones for 6400 empirical control pairs. On average, the vector Z_{jk} in this case is closer to half zeros and half ones. These types of features have a better chance of differentiating one class from another.

provide additional information about the outcome above the already-selected pairs. To accomplish this, we leverage the fact that the test statistic for $\hat{\beta}_1$ from the regression $E[Y|X] = \beta X$ is equal to the test statistic for $\hat{\gamma}_1$ from the regression $E[X|Y] = \Gamma Y$. A proof of this property is provided in **Appendix B**.

Given this property, our conditional pair selection procedure is as follows:

1. Find the pair (j, k) that maximizes the absolute test statistic for $\hat{\beta}_1$ from the regression $E[Z_{jk}|Y] = \beta Y$
2. Move the vector Z_{jk} to the right-hand side, and find the next pair (r, s) that maximizes the absolute test statistic for $\hat{\gamma}_1$ from the regression $E[Z_{rs}|Y, Z_{jk}] = \Gamma D$, where $D = \begin{bmatrix} Y Z_{jk} \end{bmatrix}$.
3. Repeat step 2 until the desired number of pairs have been found.

By flipping the regression, we need only change the right-hand side of the equation once to include the TSP chosen in the previous step. This allows for faster application of the simultaneous regression equations for all j, k at each iteration of the procedure. This process produces the set G of gene pairs that will be used in a decision tree for the prediction of Y. The key attribute of selecting features in this manner is that once Z_{jk} is chosen and added to the right-hand side, the next chosen pair, Z_{rs}, represents the maximum absolute test statistic for y_i conditional on the information provided by Z_{jk}.

3.3.4 Decision tree modeling

We use the `rpart` [81] package in R to build a decision tree using the pairs chosen through the feature selection steps described previously. We wrap a layer

of cross-validation around the entire procedure described thus far: before the empirical control step, we set up five-fold cross-validation and iterate building the tree on four-fifths of the data and predicting on the held-out one-fifth. We use cross-validation to estimate out-of-sample accuracy of the final tree model, which is built using the same procedure but using the whole data.

If the user has provided a validation dataset, then the entire training set is put through the procedure described above. If a validation dataset has not been provided, then the full training dataset is split into training and testing subsets at the outset, and the testing subset is used as a validation dataset. As a summary, we report the out-of-sample accuracy as estimated through cross-validation, the predictive accuracy of the model on the validation dataset, and use the `pROC` [71] R package to display ROC curves for both the training and validation data as a means of comparison.

3.3.5 Standardized Reporting and `tdsm` Package

We use the `rmarkdown` [10], `knitr` [90], and `knitrBootstrap` [41] packages to render an HTML report that contains code chunks as well as descriptions of the modeling process. An excerpt from an example report is shown in Figure 3.3.

The user may only provide input data to this report. Parameter choices for the number of pairs in a model or the number of empirical control pairs are fixed within the report. We have provided additional utilities should the user desire to alter these parameters. The goals of developing the `tdsm` package in this manner are to (1) be transparent about the analysis process; (2) should the user make changes to the analysis process, be transparent about what those changes are and how they affected results.

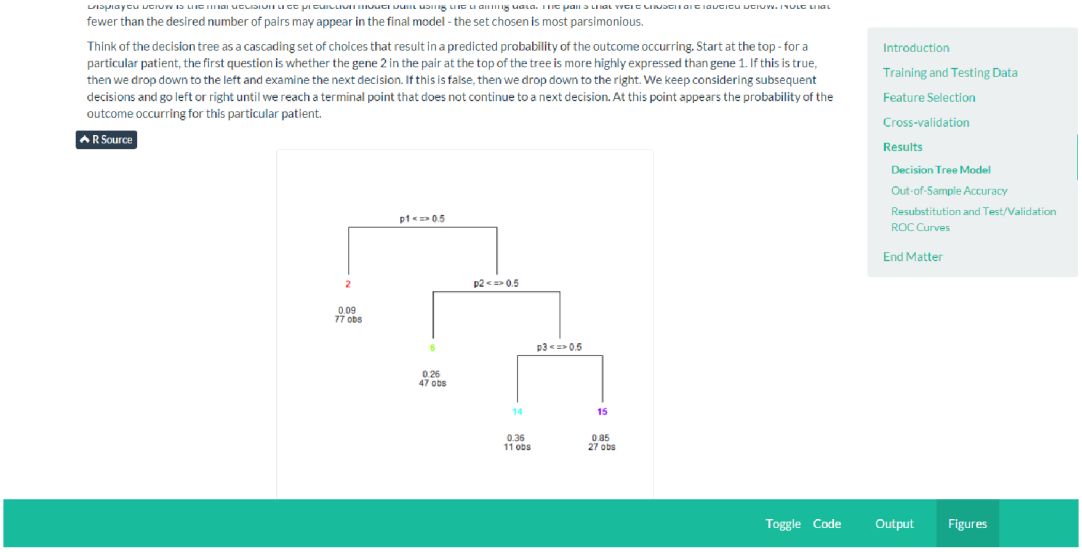

Figure 3.3: Screenshot excerpt of tspreg report This is an example of the final HTML output file produced after running the TSP-based regression analysis described herein. The report is generated from an R Markdown file styled with knitrBootstrap which takes user data as input and runs the described analysis.

Figure 3.4 shows the structure and workflow of the `tdsm` package. There are three routes a user can traverse. (In blue) the user can input data (training/testing/validation) to the `tspreg_report` function, which calls the generic .Rmd compiler `build_report`. The output is an HTML report containing a description of the analysis and all code used, as in figure 3.3. (In orange) alternatively, if the user desires to alter the analysis template in any way, they may run the `duplicate_report` command to create a new template in a directory of their choice. The user can then edit the duplicated template to their liking and reuse the `tspreg_report` command, this time providing a path to the edited template. The rest of the analysis runs along the same path as before, again producing an HTML report. The user can compare the reports from the default path (blue) and the edited path (orange) visually to note differences

due to their changes. We have also provided for file differentiation (the purple path). After the user has duplicated and edited the template, they can use the `diff_template` function to produce an HTML rendering of a `diff` command (via the `diffr` package [60], which uses the `codediff.js` Javascript library). Within this file, line differences between the original and edited templates are highlighted and documented. We also allow the user to share this HTML diff rendering online as an anonymous Github Gist with the `submit_diff` command. Our eventual intention is to collect the various changes made by different users to a particular template and determine if the default template ought to be changed to reflect common usage.

3.4 Discussion

As a proof-of-concept, we build our own gene signature for risk of breast cancer recurrence using the original MammaPrint training and validation datasets [54]. We provide the ROC curve excerpted from the full report and mark where the sensitivity and specificity of the actual MammaPrint test would fall for comparison in **Figure 3.5**. The model we developed through this procedure only uses three pairs of genes (six genes total), as compared to seventy genes used by the MammaPrint model. Both the TSP-based model and the MammaPrint model possess relatively high sensitivity and low specificity depending on the choice of a desired threshold. From the ROC curve, we see that both perform comparably on the validation data.

Examining the decision tree in **Figure 3.6**, we are more easily able to interpret how each gene pair contributes to the eventual class assignment for a

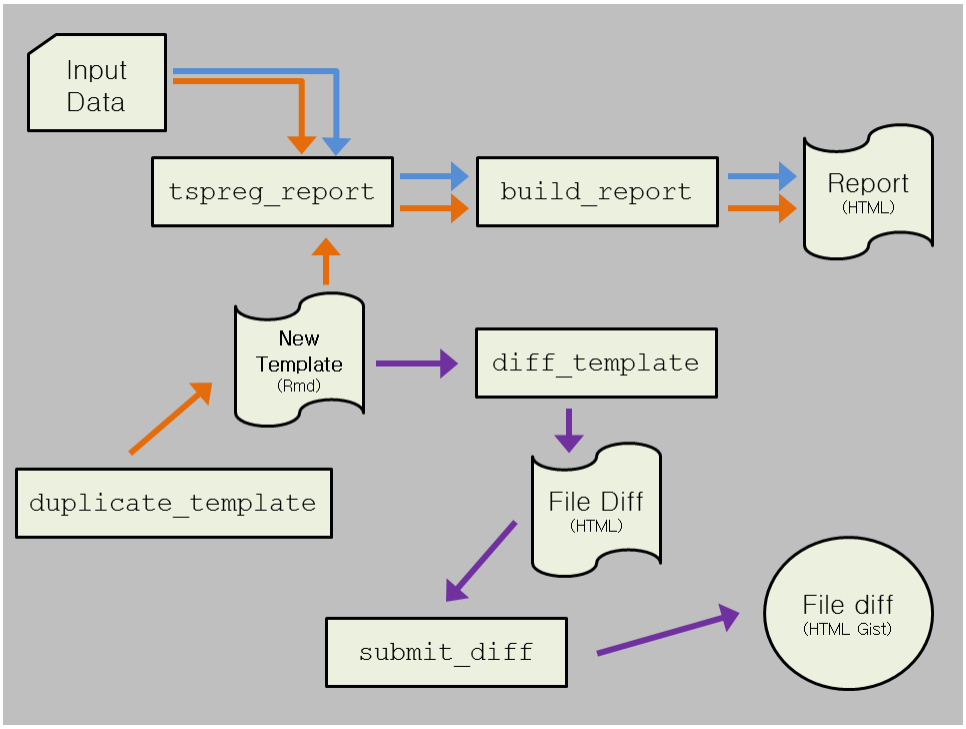

Figure 3.4: Schematic workflow of `tdsm` package. This schematic illustrates the different paths a user can take when using the `tdsm` package. In blue is the default path, where the user supplies input data to a default template and views the resulting HTML report. In orange is the alternative path, where the user chooses to edit a default template and run their input data through the edited template as opposed to the default. In this case, we recommend the addition of the purple path, where an edited template is differentiated against the default and the diff is saved to a separate HTML file. The user has the option to subsequently upload the saved diff as an anonymous Github Gist so that it may be shared and archived online.

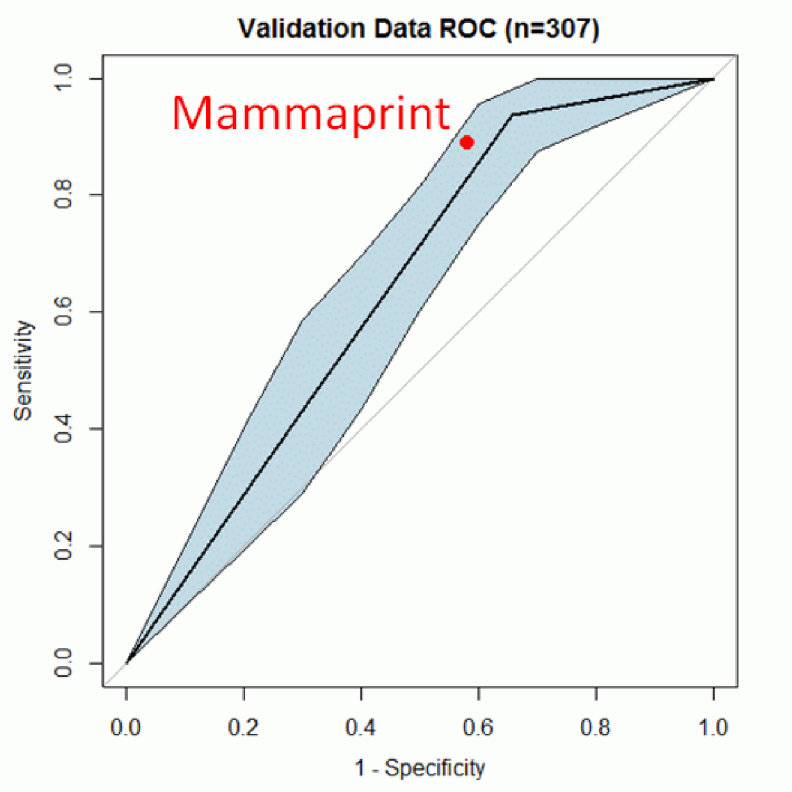

Figure 3.5: ROC curve for TSP decision tree applied to MammaPrint validation data. After building a simplified MammaPrint model on the original MammaPrint training data, we applied it to the validation data and produced an ROC curve. We plot in red the sensitivity and specificity of the MammaPrint test and note that our model performs comparably on the same validation dataset.

particular patient. One can study the genes in the most discerning pairs to determine whether their function relates to a known mechanism in breast cancer, although one of the two genes in any pair is a control gene which is unlikely to have any such relation.

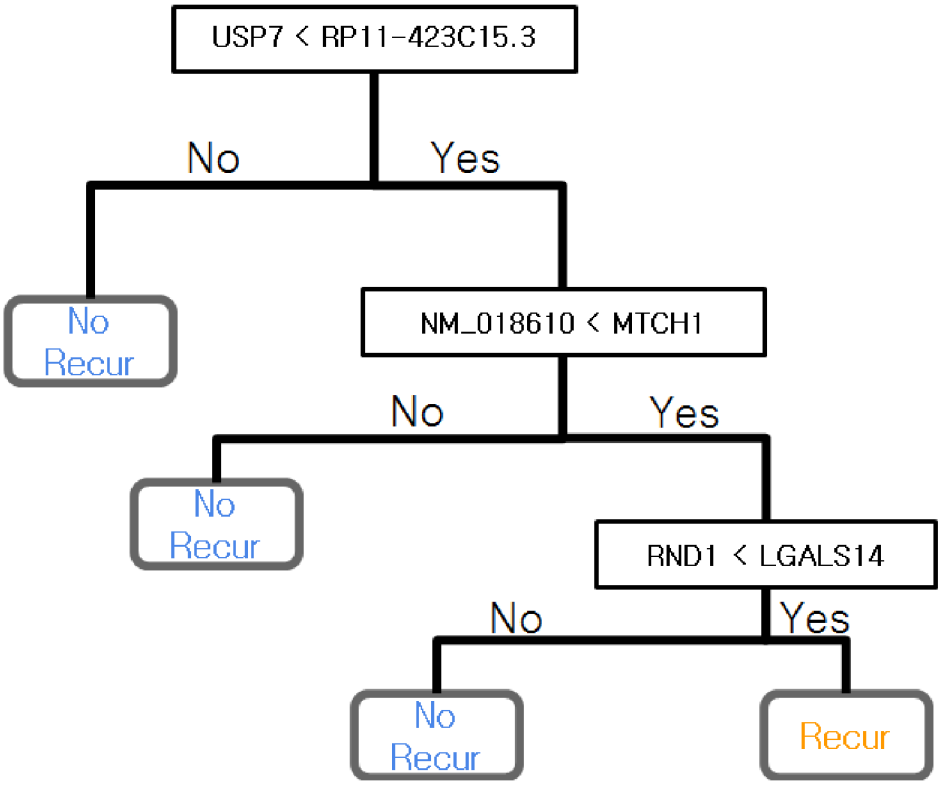

Figure 3.6: Final decision tree produced from MammaPrint training data. The final decision tree built using the MammaPrint data consisted of three TSPs (six total genes) in a cascading arrangement. This is in comparison to the seventy-gene MammaPrint model. In the decision tree model presented here, it is easier to discern how each gene pair feature is contributing to the final prediction of low or high risk for five-year recurrence of breast cancer.

3.5 Conclusion

We have described here one approach to gene signature building that avoids issues in reproducibility, both in the predictions of the model and in the model-building process. For the former, we introduce decision trees that use Top-Scoring Pairs. These are small, interpretable models that do not require data pre-processing or normalization prior to application as they operate on relative gene expression. We demonstrated the viability of this modeling procedure by recreating a simplified version of the MammaPrint risk predictor which compared favorably in sensitivity and specificity to the original.

For the latter, we have described the `tdsm` R package which contains the TSP regression procedure as its first templated analysis. Our intention with this package is to make an analysis process fully transparent and documented. By restricting user input solely to data, we produce standard reports that can be easily compared with one another. We also provide utilities in the case that the user wishes to tune the analysis to their liking. We encourage the user to differentiate their edited analysis template against the original so that the differences are documented and the resulting differences in analysis results are explicable. any normalization step when predicting on a test set.

Chapter 4

Genomic and Clinical Predictors for Improving Estimator Precision in Randomized Trials of Breast Cancer Treatments

4.1 Abstract

Background: The hope that genomic biomarkers would dramatically and immediately improve care for common, complex diseases has been tempered by slow progress in their translation beyond bioinformatics. We propose a novel use of genomic information where the goal is to improve estimator precision in a randomized trial. We analyze the potential precision gains from the popular MammaPrint genomic signature and clinical variables in simulations of randomized trials analyzed using covariate adjustment.

Methods: We apply an estimator for the average treatment effect in the trial that adjusts for prognostic baseline variables to improve precision [21]. This precision gain can be translated directly into sample size reduction and corresponding cost savings. We conduct simulation studies based on resampling

genomic and clinical data of breast cancer patients from four publicly available observational studies.

Results: Separate simulation studies were conducted based on each of the four data sets, with sample sizes ranging from 115 to 307. Adjusting only for clinical variables provided gains of -1%, 5%, 6%, 17%, compared to the unadjusted estimator. Adjusting only for the MammaPrint genomic signature provided gains of 2%, 4%, 4%, 5%. Simultaneously adjusting for clinical variables and the genomic signature provided gains of 2%, 6%, 7%, 16%. The differences between precision gains from adjusting for genomic plus clinical variables, versus only clinical variables, were -1%, 0%, 1%, 3%.

Conclusions: Adjusting only for clinical variables led to substantial precision gains (at least 5%) in three of the four data sets, with a 1% precision loss in the remaining data set. These gains were unchanged or increased when sample sizes were doubled in our simulations. The precision gains due to incorporating genomic information, beyond the gains from adjusting for clinical variables, were not substantial.

Keywords: adjustment, genomics, precision, translation

4.2 Introduction

The announcement of the Precision Medicine Initiative [23] stated that "Precision medicine's more individualized, molecular approach to cancer will enrich and modify, but not replace, the successful staples of oncology – prevention, diagnostics, some screening methods, and effective treatments – while providing a strong framework for accelerating the adoption of precision medicine in other

spheres." In the realm of genomic biomarker development, this mandate puts an explicit focus on "enrichment", i.e. how much *additional* information a new marker can provide to supplement the standard course of care. The uncertain value of genomic measurements for improving clinical practice has been a critical roadblock in the translation of genomic markers to the clinic [16], in addition to problems with reproducibility [6], interpretability [49], and cost [3]. A small number of laboratory tests based on genomic signatures have been approved for clinical use. Tests such as MammaPrint [83], Oncotype DX [61], and Prosigna [63] rely on measurement of expression for a set of genes that are associated with differential survival and severity of breast cancer cases.

It is difficult to evaluate the clinical value that these genomic signatures add beyond standard clinical factors measured for all breast cancer patients, such as age, estrogen receptor status, tumor size, and tumor grade. It is also known that tests based on genomic signatures are not part of the standard of care in many cases [25, 16]. Ongoing clinical trials are being performed to ascertain the value of some of these signatures to make adaptive treatment decisions [8]. We propose to evaluate the use of genomic signatures in a different setting by considering the prognostic value added from adjusting for a genomic signature in a randomized clinical trial of a new treatment versus control.

In a randomized trial the primary analysis typically involves estimating the average treatment effect. Adjusting for baseline variables that are prognostic for the outcome can lead to improved precision in estimating the average treatment effect at large sample sizes (i.e., asymptotically as sample size grows). [91] showed that for continuous outcomes and a linear model with main terms, the analysis of covariance (ANCOVA) estimator is guaranteed to be consistent and

as or more precise than the standard unadjusted estimator, even if the linear model is not correctly specified, i.e., the true distribution of the outcome given baseline covariates may be much more complex than the linear model used, and still the guarantee holds.

More recently, estimators with the same desirable property as the ANCOVA procedure have been extended to binary and count outcomes; see [18, 80, 73] and [36]. [21] provide a review of these recent estimators, which are designed to estimate an average treatment effect in the general setting of an observational study, where the probability of being assigned to treatment is not randomized and must be learned from the data. These estimators may also be applied to randomized trials, where their guarantees on improved precision require fewer assumptions than in an observational study since in a randomized trial the assignment probability is known (and set by design).

The above estimators all have the aforementioned consistency and precision guarantee. One difference among them is that the estimators of [91, 80]; and [21] do not require solving a non-convex (and therefore computationally challenging) optimization problem; however, the benefit of solving such a problem, as done by the estimators of [18, 73] and [36], is that they have potential for further precision gains, so there is a computation versus precision tradeoff.

The precision gains provided by adjusting for baseline variables depend on how correlated the baseline variables are with the outcome and the degree of chance imbalance in the baseline variables across the treatment groups. To the best of our knowledge, the value of such adjustment has not yet been assessed using simulations based on resampling from breast cancer patient data sets, as we do here. We resample in a way that preserves correlations between baseline

variables and the outcome in order to give a realistic assessment (as best as we can using simulations and our data sets) of the magnitude of precision gains likely to be observed in practice.

We aim to determine the prognostic value of clinical and/or genomic variables measured at baseline (pre-randomization). Of particular interest is the additional gain from adjusting for the genomic signature beyond that obtained by adjusting for standard clinical baseline variables. Our definition of precision gain in this setting equals the percent sample size reduction from using the adjusted estimator compared to the unadjusted estimator in order to attain the same power, asymptotically. Although perhaps not as groundbreaking of a result as once hoped, this approach represents a realistic attempt to assess the value of the information provided by a genomic signature.

4.3 Methods

4.3.1 Data

Microarray data used to validate the MammaPrint model [17] were gathered as described in the appendix of [55]. The MammaPrint validation data set consists of 307 breast cancer patients. **Table 4.1** summarizes the key clinical factors recorded for these patients as well as their MammaPrint risk prediction, which is a binary classification based on the risk score calculated by the MammaPrint model [83]. We dropped 11 patients whose estrogen receptor (ER) status or tumor grade were unknown and conducted our analysis using the 296 remaining patients.

We also conduct simulations based on three external breast cancer data

Characteristic	Summary
n	307
Age (years)	47.08 (7.27)
Five-Year Recurrence	
Yes	47
No	260
Tumor Size (mm)	21.48 (7.71)
Grade	
1	47
2	126
3	126
Unknown	8
ER	
+	212
−	90
Unknown	5
MammaPrint Risk Prediction	
High	194
Low	113

Table 4.1: MammaPrint validation data set. ER - estrogen receptor status, Grade - tumor severity grading (3 is most severe), Five-Year Recurrence - whether or not cancer has reappeared after five years, MammaPrint risk prediction - high or low risk for cancer recurrence. Age and Tumor Size are given as means with standard deviations in parentheses.

sets described in the Supplementary Material (**Appendix C**). These are called GSE19615, GSE11121, GSE7390, with sample sizes 115, 200, 198, respectively.

4.3.2 Statistical Method to Adjust for Baseline Covariates

We define the average treatment effect to be the difference between the population mean of the primary outcome under assignment to treatment and the population mean under assignment to control. The term "covariate adjustment" means that information from baseline variables is used to improve the precision in estimating the average treatment effect. This is done by adjusting for chance imbalances in baseline variables between treatment and control arms. Since our focus is improved precision for estimating the average treatment effect, we do not consider effects within subgroups; investigating the latter is an area for future research.

Increased precision for estimation of the average treatment effect can lead to a trial with fewer participants and shorter duration, compared to a trial with the same power that uses a less precise estimator. This is because the sample size for a trial is typically selected in order to achieve a desired power, e.g., 80% or 90%, at an alternative (e.g., the minimum, clinically meaningful effect size); using a more precise estimator leads to a smaller required sample size to achieve the power goal. More precise estimators can be used to reduce the sample size even when the average treatment effect is zero, which is the setting of our simulation study. This can be achieved by prespecifying the sample size as that which achieves a desired power at a given alternative, taking into account the percent variance reduction from using the adjusted estimator compared to

the unadjusted estimator. A more flexible approach is to use information based monitoring, where the trial runs until a preplanned information level has accrued (see, e.g., [44]). Information with respect to a given estimator, defined as the reciprocal of its variance, accrues faster for estimators with greater precision, leading to smaller sample sizes.

We assume each participant in the trial contributes a data vector $D = (W, A, Y)$, where $W = (W_1, \ldots, W_j)$ is a vector of covariates measured at baseline, A is an indicator of study arm (0 = control, 1 = treatment), and Y is a binary outcome of interest which in our case is the indicator of cancer recurrence within 5 years from baseline. We assume the trial data consist of n independent, identically distributed participant data vectors $\{D_i\}_{i=1}^n$ drawn from unknown joint distribution P on (W, A, Y). We assume a nonparametric model except that W and A are independent by randomization (called the randomization assumption), and we assume the regularity conditions in [21, Section 3.2].

The goal is to estimate the average treatment effect defined as the difference between 5 year survival probabilities comparing treatment versus control, i.e.,

$$\psi = E[Y|A = 1] - E[Y|A = 0] = P(Y = 1|A = 1) - P(Y = 1|A = 0). \quad (4.1)$$

Another possible treatment effect, which we do not consider, is the hazard ratio under a proportional hazards model. This would have the advantage that the recurrence time (rather than only the indicator Y of recurrence by 5 years) is fully used; however, a disadvantage is that inferences depend on the proportional hazards assumption being correct, and these inferences would typically be biased (even at large sample sizes) if that assumption fails to hold.

The unadjusted estimator of ψ is defined as

$$\hat{\psi}_{una} = \frac{\sum_{i=1}^{n} Y_i A_i}{\sum_{i=1}^{n} A_i} - \frac{\sum_{i=1}^{n} Y_i (1 - A_i)}{\sum_{i=1}^{n} (1 - A_i)}.$$

This estimator is consistent (i.e., converges in probability to the population average treatment effect ψ) but ignores the baseline variables W. If W is prognostic for Y then it is possible to improve precision by appropriately adjusting for W. Throughout, we do not assume that W contains information about treatment effect heterogeneity, i.e., who benefits more or less from treatment; we only use W as prognostic variables that may explain some of the variation in Y. This variation could be unrelated to treatment.

To leverage the information in W, we apply the enhanced efficiency, doubly-robust estimator of [21, Section 4.2], which is a special case of the class of estimators from [73] that is slightly modified for use in the randomized trial context. We denote this estimator by $\hat{\psi}_{adj}$. Software to compute this estimator is given in R and SAS by [21]. The R code we used is available at the link in Section 2.5.

The estimator $\hat{\psi}_{adj}$ uses parametric working models for the mean of the outcome given baseline variables and study arm. We call these working models since we do not assume they are correctly specified. The true data generating distribution may differ arbitrarily from the functional form of the model.

Computation of $\hat{\psi}_{adj}$ is accomplished via the following steps:

1. Let $\alpha = (\alpha_0, \ldots, \alpha_j)^T$. Fit the following propensity score working model for $P(A = 1|W)$: $g(W, \alpha) = \text{logit}^{-1}(\alpha_0 + \alpha_1 W_1 + \ldots + \alpha_j W_j)$ via maximum likelihood estimation and denote the estimator of α by $\hat{\alpha} = (\hat{\alpha}_0, \ldots, \hat{\alpha}_j)^T$.

2. For each arm $a \in \{0, 1\}$, define the following working model for $E(Y|A = a, W)$:
   $Q^{(a)}(W, \beta^{(a)}) = \text{logit}^{-1}\left(\beta_0^{(a)} + \beta_1^{(a)} W_1 + \ldots \beta_j^{(a)} W_j\right)$. Fit the above model at $a = 1$ using weighted logistic regression with weights $\dfrac{1}{g(W, \hat{\alpha})}$ and using only participants with $A = 1$ to obtain estimated coefficients $\hat{\beta}^{(1)} = (\hat{\beta}_0^{(1)}, \ldots, \hat{\beta}_j^{(1)})$. Define the initial estimator for $E[Y|A = 1]$ as $\hat{\mu}_1 = \dfrac{1}{n} \sum_{i=1}^{n} Q^{(1)}(W_i, \hat{\beta}^{(1)})$, where the sum is taken over all participants. The estimator $\hat{\mu}_0$ for $E[Y|A = 0]$ is obtained analogously by setting $a = 0$, replacing $A = 1$ with $A = 0$, and replacing $\dfrac{1}{g(W, \hat{\alpha})}$ by $\dfrac{1}{1 - g(W, \hat{\alpha})}$ above.

3. Define the new covariate $\mu_a(W) = Q^{(a)}(W, \hat{\beta}^{(a)}) - \hat{\mu}_a$ for each $a \in \{0, 1\}$, which uses $\hat{\mu}_a, \hat{\beta}^{(a)}$ as estimated in step 2. Fit the following augmented propensity score model for $P(A = 1|W)$: $g_{aug}(W, \alpha, \gamma) = \text{logit}^{-1}(\alpha_0 + \alpha_1 W_1 + \ldots + \alpha_j W_j + \gamma_0 \mu_0(W) + \gamma_1 \mu_1(W))$ using maximum likelihood estimation to obtain estimated coefficients $\tilde{\alpha}$ and $\tilde{\gamma} = (\tilde{\gamma}_0, \tilde{\gamma}_1)$.

4. Recompute step 2 using $g_{aug}(W, \tilde{\alpha}, \tilde{\gamma})$ in place of $g(W, \hat{\alpha})$ in the weights to obtain new estimates $\tilde{\mu}_1, \tilde{\mu}_0$. Define the adjusted estimator of the average treatment effect as $\hat{\psi}_{adj} = \tilde{\mu}_1 - \tilde{\mu}_0$.

Throughout, we assume there are no missing data and the vector (W_i, A_i, Y_i) is observed for each participant i. The models g and g_{aug} are correctly specified as long as each contains an intercept due to the randomization assumption. By design, each participant is assigned to treatment or control with probability 0.5, independent of his/her baseline variables, so $P(A = 1|W = w) = P(A = 1) =$

0.5 for all values of w. Consider the model

$$g(W, \alpha) = P(A = 1|W) = \text{logit}^{-1}(\alpha_0 + \alpha_1 W_1 + \ldots + \alpha_k W_k)$$

Setting $\alpha_1 \ldots \alpha_k = 0$ and $\alpha_0 = \text{logit}(1/2)$ yields correct specification of the model, i.e., the model at these parameter values equals the true distribution $P(A = 1|W) = P(A = 1) = 1/2$. The same holds for g_{aug}. Though the data generating distribution has A independent of W, in any given realization of the data there can be imbalances in W across arms due to chance variation.

The models $Q^{(0)}, Q^{(1)}$ will typically be misspecified if any of the baseline variables is continuous valued or has many discrete levels. An important feature of the estimator $\hat{\psi}_{adj}$ is that it is consistent regardless of whether the parametric models $Q^{(0)}, Q^{(1)}$ are correctly specified; that is, consistency holds even when the true data generating distribution $E(Y|A = a, W)$ does not have the form $Q^{(a)}(W, \beta^{(a)})$ for any β. Furthermore, the estimator $\hat{\psi}_{adj}$ is guaranteed to have asymptotic precision equal to or greater than that of the unadjusted estimator as proved by [73, 21]. However, depending on the number of baseline covariates and the sample size, the precision may be worse for the adjusted estimator compared to the unadjusted estimator; this can happen if the baseline variables are only weakly (or not at all) prognostic, there are more than a few of them, and the sample size is relatively small.

It is also possible to use the output of step 2 to construct the simpler estimator $\hat{\mu}_1 - \hat{\mu}_0$ of the average treatment effect. This estimator is called the double-robust weighted least squares estimator (DR-WLS) and is attributed to Marshall Joffe by [72]. The value of adding steps 3 and 4 is that the resulting estimator has been proved to be asymptotically as or more precise than the

unadjusted estimator [73, 21].

4.3.3 Baseline Covariates used for Adjustment

The baseline variables W used in the estimators defined above must be pre-specified. They can be any functions of measurements made prior to randomization. We define four sets of covariates that we will adjust for using the procedure described in Section 2.2:

- W_{-ER} : {Age, Tumor Size, I(Tumor Grade = 2), I(Tumor Grade = 3)}
- W_C: {Age, Tumor Size, I(Tumor Grade = 2), I(Tumor Grade = 3), ER Status}
- W_G: {MammaPrint Risk Category}
- W_{CG}: {Age, Tumor Size, I(Tumor Grade = 2), I(Tumor Grade = 3), ER Status, MammaPrint Risk Category}

Here, I(Tumor Grade = 2) is an indicator of whether or not the patient's tumor is severity grade 2.

With these four sets of covariates, we are able to contrast gains in precision from different covariate sources. We compare adjusting for W_C versus W_{-ER} to determine how much adding the clinical covariate ER status to other clinical covariates improves precision. We also compare the prognostic value of the genomic predictor plus clinical covariates (W_{CG}) versus clinical covariates alone (W_C).

We consider the clinical covariates above because they reflect quantities that clinicians may commonly use to evaluate cancer-related risks and courses of therapy. The number of covariates we are adjusting for here exceeds the conservative approach recommended by [21]. They recommend 2-3 adjustment covariates at sample sizes such as ours. The potential downside to adjusting for greater numbers of covariates is that we risk non-negligible increases in estimator variance if our covariates turn out to be non-prognostic for the outcome, as shown in Section 4.4. We chose to include larger numbers of covariates here in order to compare the added value of MammaPrint above the prognostic value of the full set of relevant clinical covariates available in our data sets.

4.3.4 Simulations

We conducted a simulation study with the goal of comparing the variance of the unadjusted and adjusted estimators to determine how much precision we may gain from adjusting for clinical and genomic covariates. For each of the four data sets described in Section 4.3.1 and in the supplement, we construct a data generating distribution that mimics the observed correlation between baseline variables and outcomes.

To preserve the relationship between outcome and potentially prognostic covariates from the original data set, we resample participants with replacement and create a new sample of the size of our data set (296 for the MammaPrint validation data) for each simulated trial; we record (W, Y) for each resampled participant. This maintains the empirical joint distribution of (W, Y), preserving the correlation of these variables. In each simulated trial, the study arm

assignment A of each participant is a random draw from the Bernoulli distribution with probability 1/2 of being 0 or 1, independent of (W, Y). The population average treatment effect defined in (4.1) corresponding to the above data generating distribution is therefore $\psi = 0$.

The reason we do not simply resample patient data vectors (W, A, Y) with replacement from a given data set is that the resulting data generating distribution would not have treatment A independent of baseline variables W (as in a randomized trial). This is because our data sets are from observational studies, as opposed to randomized trials. Though it would be preferable to use data from randomized trials, we were not able to obtain data from any such trials that also recorded the MammaPrint predictor at baseline. Observational studies still can provide a rough approximation to the magnitude of potential precision gains from covariate adjustment, since these gains are directly related to the variance of Y explained by W [21].

For each data generating distribution described above, we construct J =100,000 simulated trial data sets, each of sample size equal to the original data set (excluding patients with missing data). Using the j^{th} simulated data set, we compute the unadjusted estimator $\hat{\psi}^j_{una}$ and the adjusted estimator $\hat{\psi}^j_{adj}$ using each of the covariate sets $W_{-ER}, W_C, W_G, W_{CG}$. We then approximate the bias and variance of each of these estimators based on its values over the 100,000 simulated trials. Since $\psi = 0$, the bias B of an estimator $\hat{\psi}$ is $E(\hat{\psi}) - \psi = E(\hat{\psi})$, which is approximated by the average of $\hat{\psi}$ over the 100,000 simulated trials we conducted. We similarly approximate the variance of each estimator. For the unadjusted estimator, the approxmiate bias and variance based on our simulation study are denoted by $B_{una} = \frac{1}{J}\sum_{j=1}^{J} \hat{\psi}^j_{una}$ and $\sigma^2_{una} = \frac{1}{J-1}\sum_{j=1}^{J}(\hat{\psi}^j_{una} - B_{una})^2$,

respectively. The bias and variance approximations for the adjusted estimator $\hat{\psi}_{adj}$ are denoted similarly: $B_{adj} = \frac{1}{J}\sum_{j=1}^{J}\hat{\psi}^j_{adj}$, $\sigma^2_{adj} = \frac{1}{J-1}\sum_{j=1}^{J}(\hat{\psi}^j_{adj} - B_{adj})^2$. For conciseness, we refer the these approximations as the bias and variance of the corresponding estimator, rather than writing "approximate bias" and "approximate variance".

We define the (percent) precision gain due to the adjusted estimator in comparison to the unadjusted estimator, as approximated by simulation, as $G_{adj} = \dfrac{\sigma^2_{una} - \sigma^2_{adj}}{\sigma^2_{una}}$ x 100%. The precision gain equals, asymptotically (as sample size goes to infinity), the percent reduction in sample size to achieve a desired power at a local alternative comparing the adjusted versus unadjusted estimator. It equals $1-1/RE$, where RE is the asymptotic relative efficiency. Negative values of G_{adj} correspond to efficiency losses, which can occur if baseline variables are only weakly (or not at all) prognostic for the outcome. Asymptotically (as sample size goes to infinity), G_{adj} converges to a nonnegative value, which represents zero or positive precision gain, as proved by [73, 21].

Simulations were conducted via the `BatchJobs` R package [13], which allows for an interface between R and a cluster queuing system. We parallelized such that 1000 simulated data sets were constructed concurrently by each of 100 processors on a Sun Grid Engine (SGE) cluster, which sped up the computation of our approximations.

We also conducted simulation studies as above except where the sample size in each simulated trial is double that of the original data set. In all of our simulation studies, each simulated participant's data is an independent, identically distributed draw from a joint distribution P (which depends on the data set being resampled from) on (W, A, Y). Therefore, even though we are

resampling (with replacement) double the sample size n from the original data set, the effective sample size is $2n$ (i.e., each estimator's variance is roughly cut in half compared to its variance at the original sample size.) To illustrate this point, consider the analogy of drawing n independent, identically distributed realizations $Y_1, \ldots, Y_n$ from a Bernoulli distribution with true probability 1/4 of being 1. Though this is equivalent to resampling n times with replacement from the four person data set $\{0, 0, 0, 1\}$ (with equal chance of each), each draw is independent and the effective sample size equals the number of draws n. The precision gains from adjustment are expected to be similar or slightly greater than when the original sample sizes are used, since at larger sample sizes there is less variability in the estimated coefficients $\hat{\beta}^{(a)}$ in the working model fits $Q^{(a)}(W, \hat{\beta}^{(a)})$ used in $\hat{\psi}_{adj}$.

4.3.5 Reproducibility

Our analyses are reproducible. Code, data files, and supplementary results are available at

https://github.com/leekgroup/genesigprecision

4.4 Results

Table 4.2 presents variances for each estimator and the precision gain G_{adj}, using different sets of baseline covariates, for the MammaPrint validation data set and the data sets GSE19615, GSE11121, GSE7390. All precision gains G_{adj} are rounded to the nearest percent.

Consider the left half of **Table 4.2**, which corresponds to simulated trials having the same sample size as the corresponding data set. Adjusting only for

	Original Sample Size			Double Sample Size		
Covariate Set	σ^2_{una}	σ^2_{adj}	G_{adj}	σ^2_{una}	σ^2_{adj}	G_{adj}
MammaPrint data set						
W_{-ER}	0.0018	0.0017	4%	0.00089	0.00084	6%
W_C	0.0018	0.0017	5%	0.00089	0.00083	6%
W_G	0.0018	0.0017	5%	0.00089	0.00084	5%
W_{CG}	0.0018	0.0017	6%	0.00089	0.00082	7%
GSE19615 data set						
W_{-ER}	0.0088	0.0078	11%	0.0044	0.0037	14%
W_C	0.0088	0.0073	17%	0.0044	0.0035	21%
W_G	0.0088	0.0084	4%	0.0044	0.0042	4%
W_{CG}	0.0088	0.0074	16%	0.0044	0.0035	21%
GSE11121 data set						
W_{-ER}	0.0036	0.0034	7%	0.0018	0.0016	9%
W_C	0.0036	0.0034	6%	0.0018	0.0017	9%
W_G	0.0036	0.0036	2%	0.0018	0.0018	2%
W_{CG}	0.0036	0.0034	7%	0.0018	0.0016	9%
GSE7390 data set						
W_{-ER}	0.0045	0.0045	-1%	0.0022	0.0022	1%
W_C	0.0045	0.0045	-1%	0.0022	0.0022	1%
W_G	0.0045	0.0043	4%	0.0022	0.0022	4%
W_{CG}	0.0045	0.0044	2%	0.0022	0.0021	5%

Table 4.2: Precision gains due to adjustment for different sets of baseline covariates

clinical variables (W_C) provided precision gains G_{adj} of -1%, 5%, 6%, 17% (from smallest to largest), compared to the unadjusted estimator, across the four data sets. Adjusting only for the MammaPrint genomic signature (W_G) provided gains of 2%, 4%, 4%, 5%. Simultaneously adjusting for clinical variables and the genomic signature (W_{CG}) provided gains of 2%, 6%, 7%, 16%.

Each of the above precision gains G_{adj} was unchanged or increased when each simulated trial has double the sample size as the corresponding data set (right half of **Table 4.2**). This is to be expected, as described above. For each estimator, covariate set, and data set, the variance at double the sample size was approximately half of the corresponding variance at the original sample size, as expected.

The additional gain due to the genomic predictor is defined as the difference between the precision gain from W_{CG} versus W_C. First, consider the left half of **Table 4.2**, where each simulated trial has the same sample size as the corresponding data set. In simulations based on the MammaPrint validation data, the genomic predictor provided an additional gain of 1% above using all clinical factors. In two of the other data sets, the additional gains due to the MammaPrint predictor were 0% (GSE11121) and 3% (GSE7390). Using a third such data set, GSE19615, adjusting for the MammaPrint prediction in addition to the clinical covariates decreased precision by 1% compared to adjustment for clinical covariates alone. Such losses in precision can occur when adjusting for a variable that is only weakly prognostic (or not prognostic) for the outcome. The additional gains due to the genomic predictor were 0%, 0%, 1%, 4% when sample sizes in the simulations were doubled (right half of **Table 4.2**).

We also examined the additional gains due to ER status, defined as the

difference between the precision gains from W_C versus W_{-ER}. These values were -1%, 0%, 1%, 6%, for the four data sets, based on simulations at the original sample size. Qualitatively, these were similar to the magnitudes of additional gains due to the genomic predictor.

We conducted additional simulations where we generated baseline covariates independent of the outcome, in order to determine the magnitude of precision losses due to adjusting for pure noise. This quantifies the loss that would occur if one were to prespecify an analysis that adjusts for variables conjectured to be prognostic, but these variables turn out to be non-prognostic. We generated 100,000 simulated trial data sets as above, except where the data generating distribution has baseline variables W independent of Y. This was done by resampling W with replacement from its marginal distribution in the MammaPrint data set, and similarly resampling Y from its marginal distribution. The results are shown in **Table 4.3**. As expected, all combinations of covariates produce zero or negative precision gains, with greater losses when adjusting for larger covariate sets (due to more degrees of freedom in the working models). The maximum loss in precision is 3% when using the original sample sizes (left half of **Table 4.3**). This is due to the inclusion of greater than the recommended number of adjustment covariates, as described in Section 2.3. The potential losses are smaller if the sample size is larger, as shown in the right half of **Table 4.3** where the maximum loss is 1%. Larger sample sizes tend to decrease the magnitude of precision losses since asymptotically (as sample size goes to infinity), G_{adj} converges to a nonnegative value, which represents zero or positive precision gain, as proved by [73, 21]. We present additional simulation results with W generated independent of Y in the Supplementary Material where we

	Original Sample Size			Double Sample Size		
Covar. Set	σ^2_{una}	σ^2_{adj}	G_{adj}	σ^2_{una}	σ^2_{adj}	G_{adj}
W_{-ER}	0.00177	0.00181	-2%	0.00090	0.00091	-1%
W_C	0.00177	0.00182	-2%	0.00090	0.00091	-1%
W_G	0.00177	0.00178	0%	0.00090	0.00090	0%
W_{CG}	0.00177	0.00183	-3%	0.00090	0.00091	-1%

Table 4.3: Precision gains under data generating distribution with W and Y independent, based on marginal distributions from MammaPrint validation data set.

reduce the number of clinical covariates adjusted for, resulting in smaller precision losses.

The bias approximations B_{una} and B_{adj} were both quite small, with magnitudes of at most 0.0003 over the four simulation studies. We examined the distribution of the differences between $\hat{\psi}^j_{una}$ and $\hat{\psi}^j_{adj}$ over the $j = 1, \ldots, 100,000$ iterations in the simulation using the Mammaprint validation data set; the histogram of $\hat{\psi}^j_{una} - \hat{\psi}^j_{adj}$ appears in Figure 4.1, and analogous histograms for the other datasets along with a table comparing the distributions of differences across the four simulation studies are available in the supplement. For the simulation with the MammaPrint dataset, we saw an average difference of 0.00005 (standard deviation=0.0145). The 2.5% and 97.5% quantiles of $\hat{\psi}^j_{una} - \hat{\psi}^j_{adj}$ were [-0.029, 0.029]; this implies that 95% of the differences between the unadjusted and adjusted estimators had magnitudes smaller than 3%. The correlation of the two estimators was 0.94.

In general, we expect the difference between the unadjusted and adjusted treatment effect estimators to be small unless there is substantial chance imbalance between treatment and control arms that is accounted for by the adjusted estimator. In that case, we would expect the adjusted estimator to be closer to

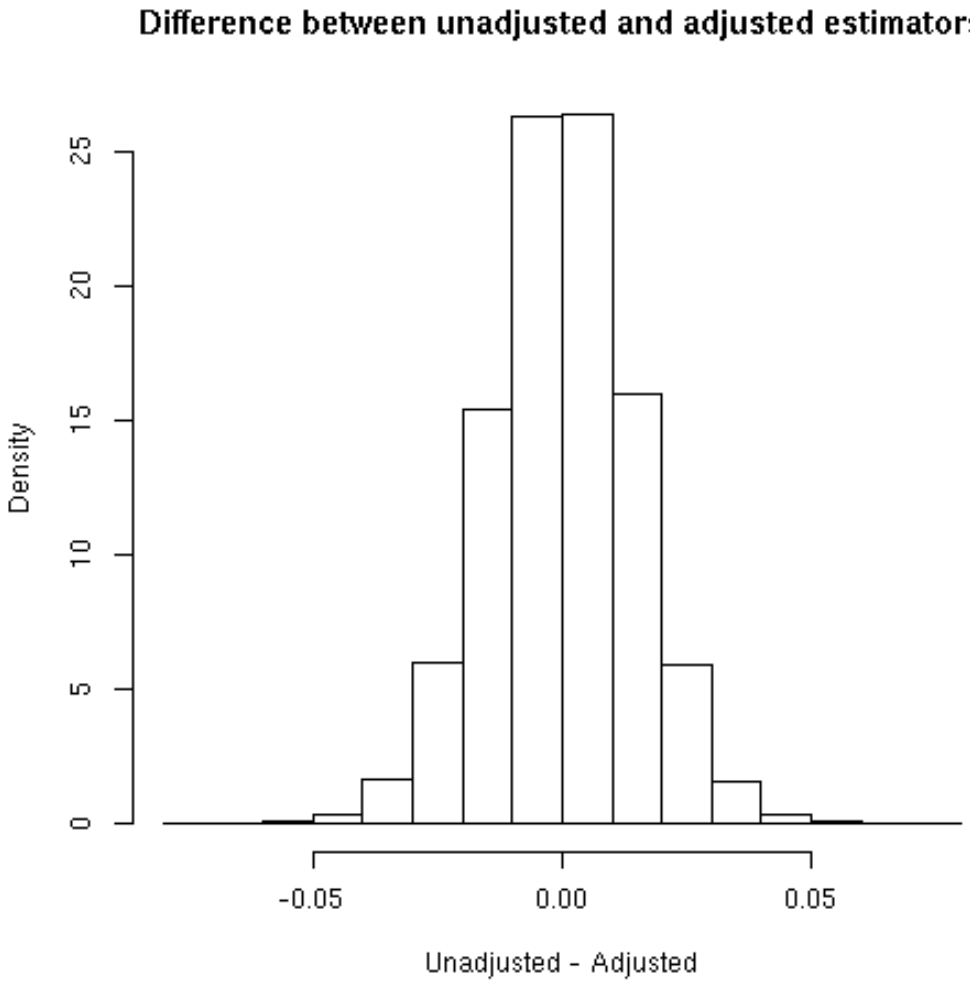

Figure 4.1: Histogram of $\hat{\psi}^j_{una} - \hat{\psi}^j_{adj}$. The histogram of differences between the unadjusted and adjusted estimators is approximately normal and is centered close to the true effect of zero (mean=0.00005, standard deviation=0.0145). The adjusted estimator is closer than the unadjusted estimator to the true effect approximately 53% of the time. For this histogram, we considered the adjusted estimator using all available baseline covariates (clinical + genomic).

the true effect. In our setting, the adjusted estimator was closer to the true effect of zero 53% of the time, suggesting a slight improvement over the unadjusted estimator.

4.5 Supplementary Material

We provide summaries of the data sets GSE19615, GSE11121, and GSE7390. We also present a table analogous to **Table 4.3** except where we reduce the maximum number of adjustment covariates to three. The result is that precision losses are smaller when covariates are generated independent of the outcome Y, compared to the precision losses in **Table 4.3**. These materials can be found in **Appendix C**.

4.6 Conclusion

Appropriately adjusting for prognostic baseline covariates has potential to improve precision in estimating the average treatment effect in randomized trials. If baseline factors are collected for patients enrolled in a study, then adjusting for them can reduce the sample size necessary to obtain a desired precision in estimation of the average treatment effect and, therefore, the cost to run the trial.

The precision gains from adjusting for clinical variables were substantial (5%, 6%, 17%) in simulation studies based on three out of four data sets we considered; the last data set led to a loss in precision of 1%. These precision gains slightly increased when sample sizes were doubled, showing that covariate adjustment can be valuable both at moderate (115 to 307) and larger sample

sizes, in the context of breast cancer treatment trials.

The additional gains from adjusting for the genomic predictor were quite small. We consider several possible explanations for this finding. First, our estimator may not have effectively extracted the additional prognostic information in the genomic marker; e.g., it may be that including interactions between the MammaPrint score and clinical variables, or using a less parametric model than logistic regression (e.g., splines), would lead to an adjusted estimator with better precision than we observed. This is difficult to evaluate, since using more flexible models could lead to overfit; this may be controllable via penalization or cross-validation, and is an area of future research. Another possible explanation is that the MammaPrint risk score is too coarse a summary measure of the 70 gene expression levels measured by the MammaPrint assay, for our purpose. The MammaPrint risk score was not designed for maximizing additional prognostic value beyond what is explained by clinical variables. It may be that a different function of the 70 gene expression levels would lead to greater precision gains, but this is beyond the scope of this paper. A third possible reason for the lackluster additional gains from the genomic predictor is that there may be little additional prognostic value in the genomic information for the outcome we considered. The MammaPrint score in the validation set examined here was 89% sensitive to high risk-of-recurrence patients, 42% specific to low risk-of-recurrence [55], but these measures ignore the variation that can be explained by clinical variables.

The additional gain due to the genomic predictor was roughly similar to the additional gain from including ER status over other clinical covariates. ER status may lack prognostic power if ER positive participants are treated with

adjuvant tamoxifen [20]. Similarly, it is possible that the MammaPrint score influenced treatment decisions, which could lead to decreased prognostic value.

A limitation of our approach is that we used data from observational studies, rather than from randomized trials. If the prognostic value of baseline variables is similar in a randomized trial setting, then our results may shed light on the order of magnitude of precision gains that can be achieved from covariate adjustment. However, if the prognostic value of baseline variables for the outcome is systematically different in a randomized trial, then our results would not apply. Future work involves applying our simulation approach to randomized trial data sets. Another limitation of our approach is that we ignore censoring due to loss to follow up. It is possible to incorporate censoring into our estimator, under a missing at random assumption, but this is an area for future work.

Our focus was on the prognostic value of different variables, that is, the ability of these variables to explain variation in the outcome (5 year recurrence). In contrast, the more ambitious goal of personalized medicine is to find predictive variables, i.e., variables that discriminate between those who are likely to benefit from a specific treatment or not. Being prognostic is not a prerequisite for being predictive, e.g., as in the case of ER status. However, the MammaPrint score having little prognostic value beyond the variation explained by clinical covariates indicates that its utility for covariate adjustment is limited.

Chapter 5

A Glass Half-Full Interpretation of Replicability in Psychological Science

5.1 Abstract

A recent study of the replicability of key psychological findings is a major contribution toward understanding the human side of the scientific process. Despite the careful and nuanced analysis reported in the paper, mass, social, and scientific media adhered to the simple narrative that only 36% of the studies replicated their original results. Here we show that 77% of the replication effect sizes reported were within a 95% prediction interval based on the original effect size. In this light, the results of *Reproducibility Project: Psychology* can be viewed as a positive result for the scientific process.

5.2 Introduction

It is natural to hope that when two scientific experiments are conducted in the same way, they will lead to identical conclusions. This is the intuition behind the

recent tour-de-force replication of 100 psychological studies by the Open Science Collaboration, *Reproducibility Project: Psychology* [22]. At incredible expense and with painstaking effort, the researchers attempted to replicate the exact conditions for each experiment, collect the data, and analyze them identically to the original study.

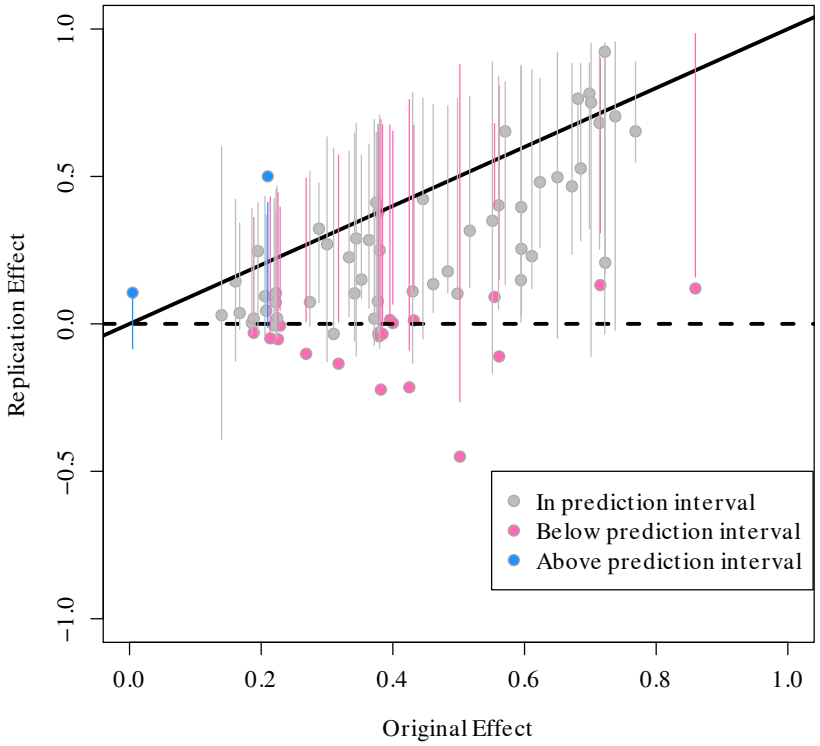

Figure 5.1: 95% prediction intervals suggest most replication effects fall in the expected range A plot of original effects on the correlation scale (x-axis) and replication effects (y-axis). Each vertical line is the 95% prediction interval based on the original effect size. Replication effects could either be below (pink), inside (grey), or above (blue) the 95% prediction interval.

The original analysis considered both subjective and quantitative measures

of whether the results of the original study were replicated in each case. They compared average effect sizes, compared effect sizes to confidence intervals, and measured subjective and qualitative assessments of replication. Despite the measured tone of the manuscript, the resulting mass, social, and scientific media coverage of the paper fixated on a statement that only 36% of the studies replicated the original result [66].

Although we may hope that a properly replicated study will provide the same result as the original, foundational statistical principles suggest that this may not be the case. The *Reproducibility Project: Psychology* study coincided with extensive discussion on what it means for a study to be reproducible and how to account for different sources of variation when replicating [45]. Stanley and Spence [78] showed through simulation how sampling and measurement variation interplay with the size and reliability of an effect to produce wide distributions of replication effect sizes. These examinations were accompanied by discussions of adequate study power [56, 58], sample size [29, 74], and how meta-analysis may address the consequences of inadequate power or sample size [15]. Anderson and Maxwell [2] furthered these concepts by categorizing the different goals of replicating a study and recommending appropriate analyses and equivalence tests specific to each goal. In sum, the sources of variability that make replicating the result of a particular study so difficult were well-documented when the *Reproducibility Project: Psychology* study was underway.

5.3 Defining and Quantifying Replication Using P-values

Despite the nuanced understanding of the factors that affect reproducibility, the publicized 36% figure refers only to the percentage of study pairs that reported a statistically significant ($P < 0.05$) result in both the original and replication studies. The relatively low number of results that were statistically significant in both studies was the focus of extreme headlines like "Over half of psychology studies fail reproducibility test." [7] and played into the prevailing narrative that science is in crisis [30].

The most widely disseminated report from this paper is based on a misinterpretation of reproducibility and replicability. Reproducibility is defined informally as the ability to recompute data analytic results conditional on an observed data set and knowledge of the statistical pipeline used to calculate them [67, 68]. The expectation for a study to be reproducible is that the exact same numbers will be produced from the same code and data every time. Replicability of a study is the chance that a new experiment targeting the same scientific question will produce a consistent result [4, 42]. When a study is replicated, it is not expected that the same numbers will result for a host of reasons including both natural variability and changes in the sample population, methods, or analysis techniques [48].

We therefore do not expect to get the same answer even if a perfect replication is performed. Defining replication as consecutive results with $P < 0.05$ squares with the intuitive idea that replication studies should arrive at similar conclusions. So it makes sense that despite the many reported metrics in the

original paper, the media has chosen to focus on this number. However, this definition is flawed since there is variation in both the original study and in the replication study, as has been much-studied in the psychology community to date. Even if you performed 10,000 perfect studies and 10,000 perfect replications of those studies, you would expect the number of times both P-values to be less than 0.05 to vary.

We conducted a small simulation based on the effect sizes presented in the original article. In the original study, the authors applied transformations to 73 of the 100 studies whose effects were reported via test statistics other than the correlation coefficient (e.g. t-statistics, F-statistics). We simulated 10,000 perfect replications of these 73 studies based on one degree of freedom tests. Each of these 10,000 simulations represents a perfect version of the Reproducibility Project with no errors. In each case, we calculated the percentage of P-values less than 0.05. The percentage of P-values less than 0.05 ranged from 73% to 91% (1^{st} to 3^{rd} quartile; high: 100%; low: 6%) with a high degree of variability (**Figure 5.2**).

5.4 Prediction Intervals

Sampling variation alone may contribute to "un-replicated" results if you define replication by a P-value cutoff. We instead consider a more direct approach by asking the question: "What effect would we expect to see in the replication study once we have seen the original effect?" This expectation depends on many variables about how the experiments are performed [35]. Here we

assume the replication experiment is indeed a true replication - a not unreasonable assumption in light of the effort expended to replicate these experiments accurately.

One statistical quantity that incorporates what we can reasonably expect from subsequent samples is the prediction interval. A traditional 95% confidence interval describes our uncertainty about a population parameter of interest. We may see an odds ratio reported in a paper as 1.6 [1.2, 2.0]. Here, 1.6 is our best estimate of the true population odds ratio based on the observed data. The range [1.2, 2.0] is our 95% confidence interval constructed from this study. If we were able to observe 100 samples and construct a 95% confidence interval for each sample, 95 of the 100 would contain the true population odds ratio.

A prediction interval makes an analogous claim about an individual future observation given what we have already observed. In our context, given the observed original correlation and some distributional assumptions (described in detail in the Methods section), we can construct a 95% prediction interval and state that if we were to replicate the exact same study 100 times, 95 of our observed replication correlations will fall within the corresponding prediction interval.

A crucial characteristic of the prediction interval which makes it a suitable tool for our purposes is that it takes the variability in the observed data as well as the future data point into account. This is shown explicitly through calculation in the Methods section, but the basic concept is that constructing a prediction interval relies on computing a contrast between a summary of the observed data (such as the mean, $\bar{X}$) and the theoretical "next" observation, X_{n+1}. Then the variance of the distribution of $\bar{X} - X_{n+1}$ will depend on the

variance of both $\bar{X}$ and X_{n+1}, as will the subsequent interval calculated from that distribution.

5.5 Using Prediction Intervals to Assess Replication

Assuming the replication is true and using the derived correlations from the original manuscript, we applied Fisher's z-transformation [27] to calculate a pointwise 95% prediction interval for the replication effect size given the original effect. The 95% prediction interval is $\hat{r}_{orig} \pm z_{0.975}\sqrt{\frac{1}{n_{orig}-3} + \frac{1}{n_{rep}-3}}$, where $\hat{r}_{orig}$ is the correlation estimate in the original study; n_{orig}, n_{rep} are the sample sizes in the original and replication studies; and $z_{0.975}$ is the 97.5% quantile of the normal distribution (Methods). The prediction interval accounts for variation in both the original study and in the replication study through the sample sizes incorporated in the expression of the standard error.

We observe that for the 92 studies where a replication correlation effect size could be calculated, 69 (or 75%) were covered by the 95% prediction interval based on the original correlation effect size (**Figure 5.1**). In two cases, the replication effect was actually larger than the upper bound of the 95% prediction interval. Considering the asymmetric nature of the comparison, one might consider these effects as having "replicated with effect clear". We then estimate that 71/92 (or 77%) of replication effects are in or above the 95% prediction interval based on the original effect. Some of the effects that changed signs upon replication still fell within the 95% prediction intervals calculated based on the original effects. This in unsurprising in light of the relatively modest sample sizes and effects in both the original and replication studies (**Figure 5.3**).

We note here that of the 69 replication effect sizes that were covered by the 95% prediction interval, two replications showed a slightly negative correlation (-0.005, -0.034) as compared to a positive correlation in the original study (0.22, 0.31, respectively). In the first study, the original and replication sample sizes were 110 and 222; in the second study, they were 53 and 72. We would classify these two studies as "replicated with ambiguous effect" as opposed to "replicated with effect clear" due to the change in direction of the effect, although both are very close to zero. All other negative replication effects did not fall into the 95% prediction intervals, and hence were considered "did not replicate".

We also considered the 73 studies the author's reported to be based on one degree of freedom tests. In 51 of these 73 studies (70%), the replication effect was within the 95% prediction interval. The same two cases where the replication effect exceeded the 95% prediction interval were in this set leaving us with an estimate of 53/73 (73%) of these studies had replication effects consistent with the original effects.

Based on the theory of the prediction interval we expect about 2.5% of the replication effects to be above and 2.5% of the replication effects to be below the prediction interval bounds. Since about 23% were below the bounds, this suggests that not all effects replicate or that there were important sources of heterogeneity between the studies that were not accounted for. The key message is that replication data—even for studies that should replicate—is subject to natural sampling variation in addition to a host of other confounding factors.

It is notable that almost all of the replication study effect sizes were smaller than the original study effect sizes, whether or not they fell inside the 95% prediction interval. In the original set of 92 studies, of those where the replication

effect falls within the 95% prediction interval (69 studies), 55/69 (80%) had a replication effect size that was smaller than the original effect size. This speaks to the notion that there are likely a host of biases that pervade the original study, pertaining mostly to the desire of reporting a small effect that is statistically significant [31]. In this sense, our analysis complements the finding of the Open Science Collaboration while simultaneously providing some additional perspective on the expectation of replicability.

5.6 Conclusion

We need a new definition for replication that acknowledges variation in both the original study and in the replication study. Specifically, a study replicates if the data collected from the replication are drawn from the same distribution as the data from the original experiment. To definitively evaluate replication we will need multiple independent replications of the same study. This view is consistent with the long-standing idea that a claim will only be settled by a scientific process rather than a single definitive scientific paper. We support Registered Replication Reports [76] and other such policies that incentivize researcher contribution to these efforts.

The *Reproducibility Project: Psychology* study highlights the fact that effects may be exaggerated and that replicating a study perfectly is challenging. We were caught off guard by the immediate and strong sentiment that psychology and other sciences may be in crisis [30]. Our first reaction to Figure 3 from the original manuscript was pleasant surprise. The fact that many effects fall within the predicted ranges despite the long interval between original and replication

study, the complicated nature of some of the experiments, and the differences in populations and investigators performing the studies is a reason for optimism about the scientific process. It is also in line with estimates we have previously made about the rate of false discoveries in the medical literature [43]. While there is a work to be done, the glass may not be quite as empty as the prevailing narrative would suggest.

We stress that the approach outlined here of is easily applied when the result of interest in a study can be summarized by one value upon which we can ascribe distributional assumptions. In reality, most scientific studies are quite a bit more complex, dealing in multiple stimuli [87], adaptation over time and circumstance [12], and complicated data sources [19], just to name very few. Our suggestion of 95% prediction intervals to help assess replication is meant to establish a conceptual framework and motivate researchers to simply begin considering what is a reasonable expectation for a replicated effect. Extending these concepts to modern study designs is the next step in auditing the conduct of scientific research.

5.7 Methods

5.7.1 Calculating a 95% Prediction Interval

Comparing 95% Confidence Interval Calculation to 95% Prediction Interval Calculation

Suppose we observe data $X_1, \ldots, X_n$ from a normal distribution with mean μ and variance σ^2, with σ^2 known. Then by the Central Limit Theorem, $\bar{X} \sim N(\mu, \sigma^2/n)$ and $Z = \dfrac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$. We can state that $P(-z_{0.975} < Z <$

$z_{0.975}) = 0.95$. This leads to the following arithmetic:

$$\begin{aligned} P(-z_{0.975} < Z < z_{0.975}) &= 0.95 \\ P(-z_{0.975} < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < z_{0.975}) &= 0.95 \\ P(-\bar{X} - z_{0.975}\frac{\sigma}{\sqrt{n}} < -\mu < -\bar{X} + z_{0.975}\frac{\sigma}{\sqrt{n}}) &= 0.95 \\ P(\bar{X} - z_{0.975}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{0.975}\frac{\sigma}{\sqrt{n}}) &= 0.95 \end{aligned}$$

Hence, we state the 95% confidence interval for μ as $\bar{X} \pm z_{0.975}\sigma/\sqrt{n}$.

Contrast this with the procedure with constructing a 95% prediction interval for X_{n+1}, another observation from the original normal distribution. From above, $X_{n+1} \sim N(\mu, \sigma^2)$ and $\bar{X} \sim N(\mu, \sigma^2/n)$. Since both are normally distributed, we are able to state the distribution of their difference: $\bar{X} - X_{n+1} \sim N(0, \sigma^2 + \sigma^2/n) = N(0, \sigma^2(1 + 1/n))$. This means that $\dfrac{\bar{X} - X_{n+1}}{\sigma\sqrt{1 + 1/n}} \sim N(0, 1)$. We can run through a series of similar arithmetic as above to isolate X_{n+1} and produce the 95% prediction interval for X_{n+1} as $\bar{X} \pm z_{0.975}\sigma\sqrt{1 + 1/n}$. Notice that in the calculation of the prediction interval, both the variability of the observed sample $X_1, \ldots, X_n$ as well as the variability of the new observation X_{n+1} come into play.

Note that the 95% confidence interval for the true difference (here $\mu - \mu = 0$) is $\bar{X} - X_{n+1} \pm z_{0.975}\sigma\sqrt{1 + 1/n}$, and that this is referring to a population difference of zero. Let us suppose that once this confidence interval is constructed, zero is outside and strictly below the interval. Then we would have the following equations:

$$(\bar{X} - X_{n+1}) - \sigma\sqrt{1 + 1/n} \ > 0$$

$$(\bar{X} - X_{n+1}) + \sigma\sqrt{1 + 1/n} \ > 0$$

If we add our quantity of interest for the prediction interval, X_{n+1}, to both sides, we have:

$$\bar{X} - \sigma\sqrt{1 + 1/n} \ > X_{n+1}$$

$$\bar{X} + \sigma\sqrt{1 + 1/n} \ > X_{n+1}$$

A similar argument can be made if zero is outside and strictly above the 95% confidence interval for the difference. Hence, the new observation X_{n+1} falls inside the 95% prediction interval if and only if zero, the true population difference, falls inside the 95% confidence interval for the difference.

95% Prediction Interval for Correlation Coefficients

We calculate a prediction interval based on the original r_{orig} and replication r_{rep} correlation estimates. Under normality and independence assumptions, the Fisher z-transformation provides the relationship:

$$z^f = arctanh(\hat{r}) = \frac{1}{2}\log\left(\frac{1+\hat{r}}{1-\hat{r}}\right) \sim N\left(\frac{1}{2}\left(\frac{1+\rho}{1-\rho}\right), \frac{1}{N-3}\right)$$

Assume that $\hat{r}_{orig}$ and $\hat{r}_{rep}$ are the estimates from the original and replication studies and assume they have a common value. Then, we make the conservative assumption that the original and replication experiments are independent, we can calculate:

$$\hat{z}^f_{orig} - \hat{z}^f_{rep} \sim N\left(0, \frac{1}{n_{orig}-3} + \frac{1}{n_{rep}-3}\right)$$

then letting $se_{total} = \sqrt{\frac{1}{n_{orig}-3} + \frac{1}{n_{rep}-3}}$ we know that:

$$\frac{1}{se_{total}}(\hat{z}^f_{orig} - \hat{z}^f_{rep}) \sim N(0,1)$$

so we have that:

$$\begin{aligned} P(-z_{1-\alpha/2} < \frac{1}{se_{total}}(\hat{z}^f_{orig} - \hat{z}^f_{rep}) < z_{1-\alpha/2}) &= 1-\alpha \\ P(-\hat{z}^f_{orig} - z_{1-\alpha/2}se_{total} > -\hat{z}^f_{rep} > -\hat{z}^f_{orig} + z_{1-\alpha/2}se_{total}) &= 1-\alpha \\ P(\hat{z}^f_{orig} - z_{1-\alpha/2}se_{total} < \hat{z}^f_{rep} < \hat{z}^f_{orig} + z_{1-\alpha/2}se_{total}) &= 1-\alpha \end{aligned}$$

So a $(1-\alpha)\%$ prediction interval for z^f_{rep} is $\hat{z}^f_{orig} \pm se_{total}z_{1-\alpha/2}$. We can then apply the inverse of the Fisher z-transform to obtain bounds for r_{rep} on the appropriate scale.

5.7.2 P-value Simulation

To simulate, we took all reported correlation coefficients for original studies. As described above, $arctanh(\hat{r}) \sim N\left(\frac{1}{2}\left(\frac{1+\rho}{1-\rho}\right), \frac{1}{N-3}\right)$. We set $\rho = \hat{r}_{orig}$, and simulated 100 times, for each study, from the distribution $N\left(\frac{1}{2}\left(\frac{1+\hat{r}_{orig}}{1-\hat{r}_{orig}}\right), \frac{1}{n_{rep}-3}\right)$, where n_{rep} was the sample size of each replication experiment. If n_{rep} was unavailable, we used n_{orig}, and if both were unavailable, we used the median of the original sample sizes. If $\hat{r}_{orig}$ was unavailable, we similarly used the median of the correlations coefficients for the original studies.

Once we had 10,000 realizations from the distribution of the replicate correlation coefficients, we back-calculated them into F-statistics. We used the formula from the supplement of the original paper: $r = \sqrt{\frac{F\frac{df_1}{df_2}}{F\frac{df_1}{df_2}+1}}\sqrt{\frac{1}{df_1}}$. From these 10,000 F-statistics, we were able to calculate 10,000 P-values and count up how many were < 0.05

We made two assumptions/simplifications in the course of running this simulation, as it is merely for illustrative purposes. (1) We assumed that the correlation coefficient reported for the original study represented the true, population correlation coefficient (2) we converted all simulated correlation coefficients to $F(1, df_2)$ statistics, where df_2 were the degrees of freedom from the size of the replication study. Since 70% of the original studies conducted the same analysis, we felt that this was a reasonable simplification for comparative purposes.

5.7.3 Code

Code and data to reproduce this analysis is available from:

- https://github.com/jtleek/replication_paper

- http://jtleek.com/replication_paper/code/replication_analysis.html

5.8 Supplementary Figures

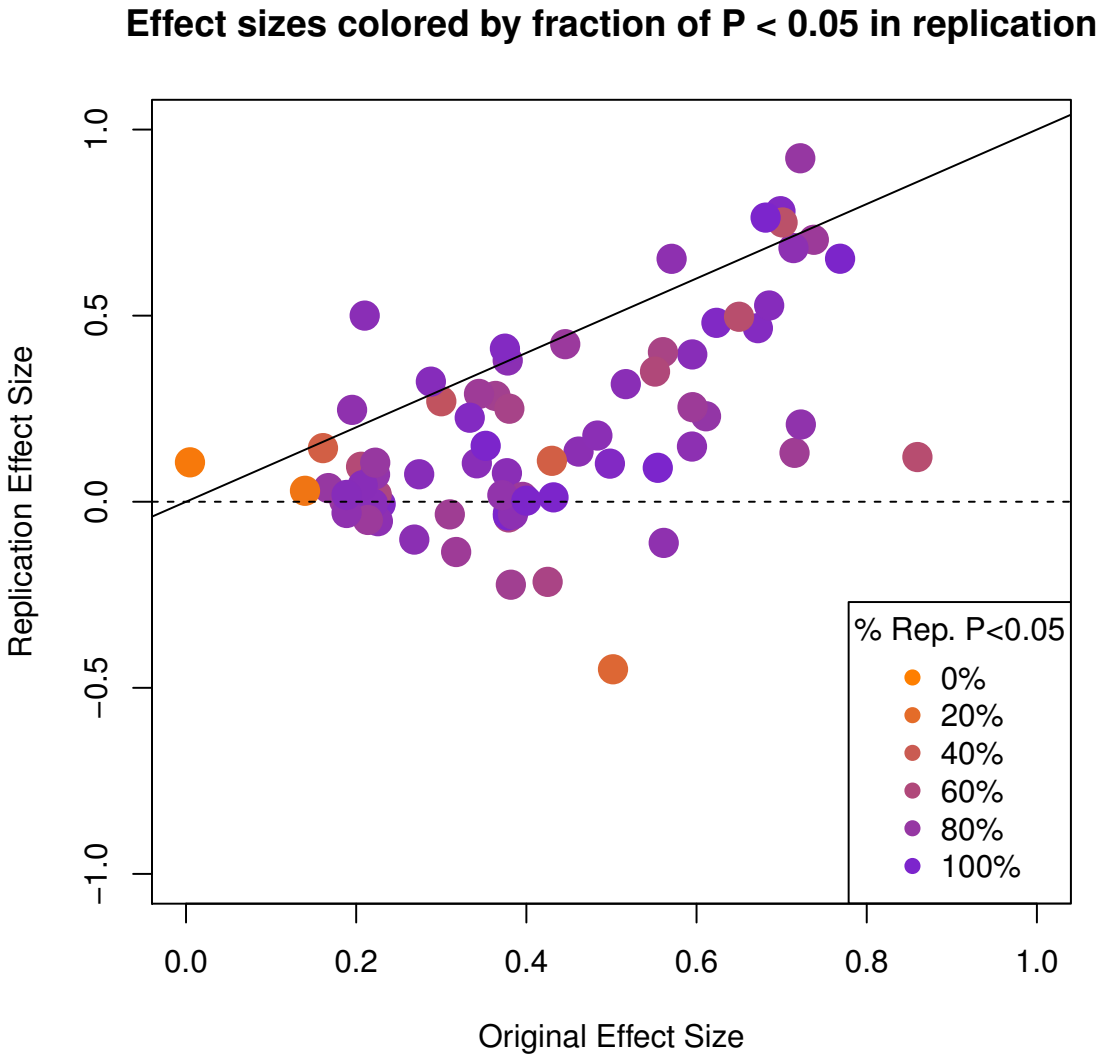

Figure 5.2: Empirical probability of replicating by effect size. We simulated 10,000 effects from a distribution that assumes the original study effect is true. These were converted to test statistics, for which P-values were calculated. We then colored each point from Figure 3 in the original paper by how many times the calculated P-value was < 0.05 out of the 10,000 simulations. This corresponds to the empirical probability of each study "replicating" by twice showing a statistically significant P-value

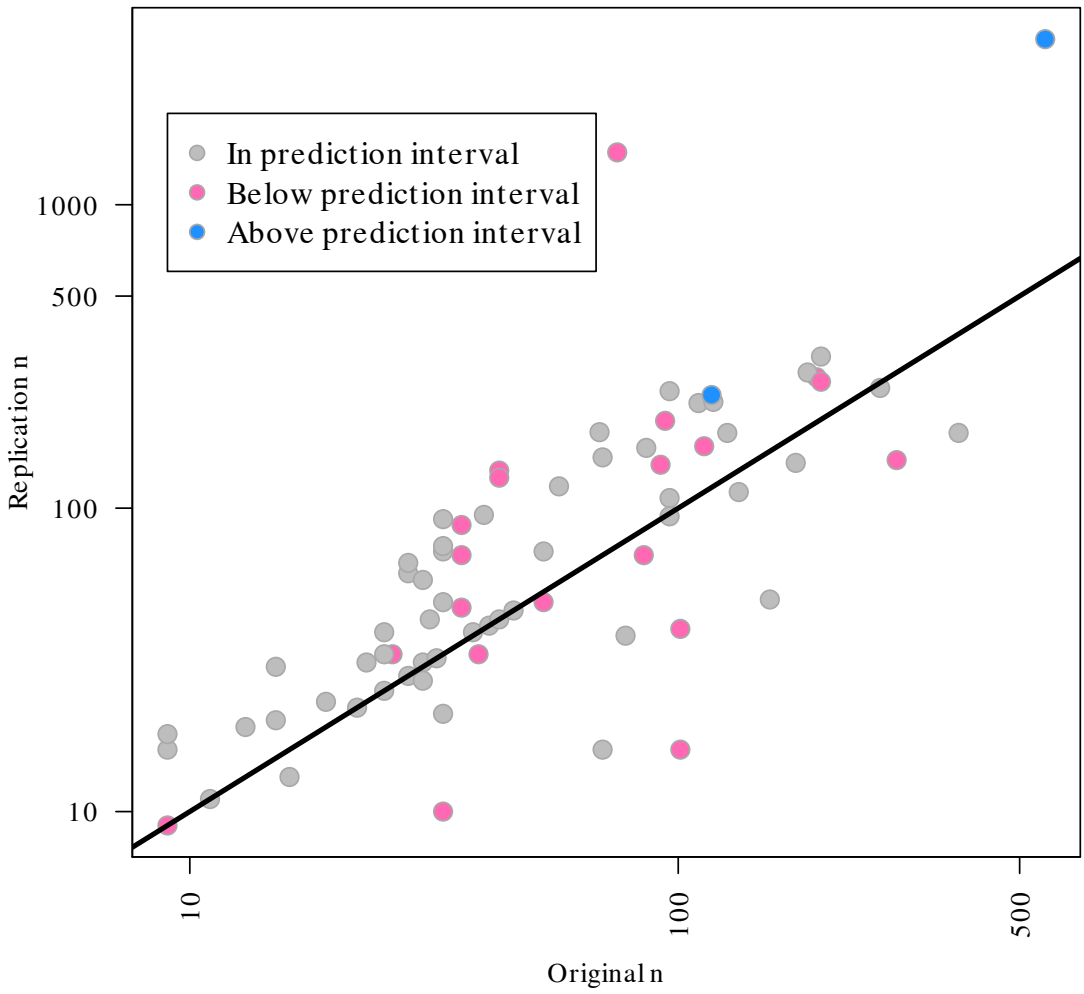

Figure 5.3: Sample sizes of studies in the Reproducibility Project colored by whether they fell in the 95% prediction interval. A plot of the original versus replication sample size colored by whether the resulting replication effect was inside (grey), above (blue) or below (pink) the 95% prediction interval based on the original effect size.

Chapter 6

Conclusion

In this compendium of work, we have proposed a set of novel approaches for building and investigating genomic signatures. All of these proposals have been made with downstream clinical impact in mind. Specifically, we examined issues in reproducibility, replicability, and additional value provided by genomic prediction.

We first described an issue in the reproducibility of predictions from existing gene signatures based on gene expression information. In introducting "test set bias", we illustrated how necessary steps in normalization and data preprocessing depend upon the size and characteristics of the test set of patients upon which we are applying the predictive model. We applied the PAM50 subtyper to a real breast cancer data set and showed that patients in subsamples of different sizes or proportions of Estrogen Receptor negative patients may receive a different subtype assignment than they did when processed with the entire data set. Since this is simply a technical context change and not a true biological change in the patient, we deem this to be problematic and suggest rank-based prediction as an alternative that can avoid this form of bias.

We then suggested a novel signature-building method that uses rank-based features (Top-Scoring Pairs) as primary predictors in a decision tree. These choices are motivated by our desire to make the entire signature-building process more transparent, interpratable, and reproducible. We described "empirical controls" and conditional feature addition via regression as filtration and wrapper steps unique to Top-Scoring Pair selection. We compared a small decision tree built from our method that relies on fewer than ten genes to the MammaPrint signature, which relies on seventy genes, and showed that the performance of our signature was comparable on the original MammaPrint validation data. To ensure reproducibility and transparency of the procedure, we described the `tdsm` R package, which provides templated data analyses with user input as the only parameter.

Supposing that we have built a genomic signature that does not succumb to "test set bias" and is well-defined and interpretable, the question remains of how much additional value a prediction from this signature would provide above what a doctor already knows when a breast cancer patient is examined in the clinic. To address this question, we proposed the use of covariate adjustment techniques in the realm of randomized clinical trials. We described a class of adjusted estimators that provide as much or more precision (small variance) when compared to the standard, unadjusted average treatment effect estimator. We leveraged the fact that we can estimate precision gain due to covariate adjustment to test different sets of covariates in a trial simulation based on real data with exogenous assigned treatment. Of specific interest was comparing the relative gain due to adjusting for a set of only clinical covariates (Age, ER Status, Tumor Size, Tumor Grade) to the same set with the MammaPrint

prediction included. We showed that an additional 1-2% gain in precision can be attributed directly to the prediction from MammaPrint.

Finally, we described the issue of replicability of a scientific study through an example in the psychology literature. We suggested 95% prediction intervals as a method to establish expectations of a plausible range of values within which a replicated finding could reasonably fall. We used this approach to provide context to a recent study on replicability in psychology and were able to assure that most of the replicated effects fell within a plausible range when the variability in both the original study and the replication were taken into account. We included this work in reference to genomic signatures to emphasize the need for multiple validation studies for confirmation of an association.

Public Health Impact

The goal of this work is to improve the standing of genomic predictors in clinical use. By assessing their reproducibility, replicability, and value, we hope that signatures produced in the future will be more reliable, consistent, and trustworthy. Assuming that the underlying relationship between genomic features and the outcome of interest is informative, we hope that these improved signatures have greater potential to be part of the standard of care for patients. For cancer patients, a genomic test may be less invasive than a tumor biopsy or other standard clinical technique, so if tests based on genomics are truly providing additional value in a reliable manner, there is great opportunity to substantially improve patient care.

Chapter 7

Appendices

7.1 Appendix A

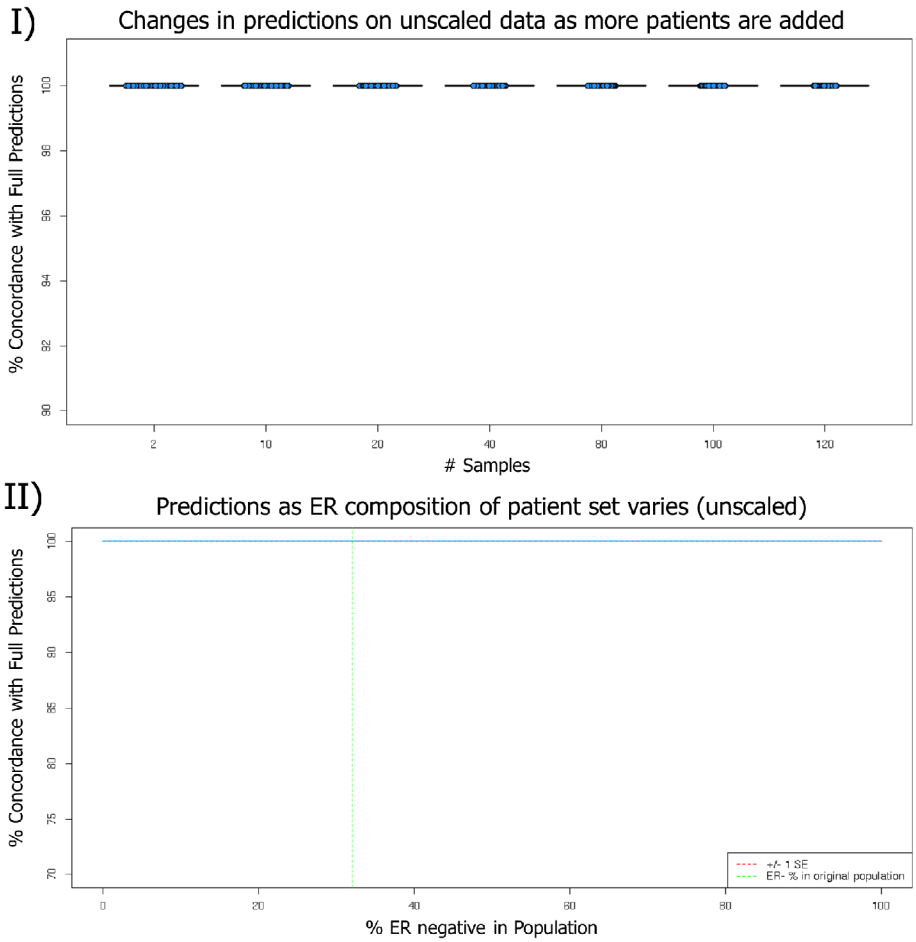

Figure 7.1: Classifications remain unchanged when rank-based classification is used We conducted the exact same simulation as in **Figure 2.2**, except we used the unscaled version of PAM50 to make predictions. We find that in this case, classifications do not change by sample size or ER status as they did when rescaling was first applied to the data. This provides empirical justification for the use of rank-based classifiers that do not need to rely on normalization and data scaling procedures to make predictions.

7.2 Appendix B

7.2.1 Proof of t-statistic equivalency when regression is flipped

Suppose we have vectors X, Y, Z_1, Z_2 of length n x 1. Let $Z = \begin{bmatrix} 1 & Z_1 & Z_2 \end{bmatrix}$, $D = \begin{bmatrix} X & Z \end{bmatrix}$, $\beta = \begin{bmatrix} \beta_1 & \beta_0 & \beta_2 & \beta_3 \end{bmatrix}$, $H = I - Z(Z'Z)^{-1}Z'$. We organize β as stated so that β_1 is the multiple regression coefficient for X. We establish a partial least squares regression as follows:

$$\begin{aligned} Y &= D\beta' + \epsilon \\ HY &= HD\beta' + H\epsilon \\ e_{Y|Z} &= e_{X|Z}\beta_1 + \epsilon^* \end{aligned}$$

Here, $e_{Y|Z}$ is the vector of residuals from regressing Y on Z. We will then use the least squares solution for β_1 in terms of residuals, and recall the definition that $cor(a,b) = cov(a,b)sd(a)sd(b)$:

$$\begin{aligned} \hat{\beta}_1 &= (e'_{X|Z}e_{X|Z})^{-1}e'_{X|Z}e_{Y|Z} \\ &= \frac{cov(e_{X|Z}, e_{Y|Z})}{var(e_{X|Z})} \\ &= \frac{cor(e_{X|Z}, e_{Y|Z})sd(e_{X|Z})sd(e_{Y|Z})}{sd(e_{X|Z})^2} \\ &= \rho_{X,Y|Z}\frac{sd(e_{Y|Z})}{sd(e_{X|Z})} \end{aligned}$$

where $\rho_{X,Y|Z}$ is the parital correlation coefficient for X and Y given Z, by definition the correlation of the residuals above.

Next, recall that $\hat{var}(\hat{\beta}) = s^2(D'D)^{-1}$. In the residual equation, this simplifies to

$$\frac{\frac{1}{n-p}\sum(e_{Y|Z} - \hat{e}_{Y|Z})^2}{(n-1)var(e_{X|Z})} = \frac{\sum(e_{Y|Z} - \hat{e}_{Y|Z})^2}{(n-p)(n-1)var(e_{X|Z})}$$

In the above equation, we must set p to the number of parameters in the original multiple regression, $Y = D\beta'$, to ensure proper calculation of the variance.

Combining the first two steps, we have the expression for the t-statistic for β_1, denoted t_X, in terms of the partial regression equation:

$$t_X = \frac{\rho_{X,Y|Z}\dfrac{sd(e_{Y|Z})}{sd(e_{X|Z})}}{\sqrt{\dfrac{\sum(e_{Y|Z} - \hat{e}_{Y|Z})^2}{(n-p)(n-1)var(e_{X|Z})}}}$$

We manipulate this expression as follows:

$$\begin{aligned}
t_X &= \frac{\rho_{X,Y|Z}\dfrac{sd(e_{Y|Z})}{sd(e_{X|Z})}}{\sqrt{\dfrac{\sum(e_{Y|Z} - \hat{e}_{Y|Z})^2}{(n-p)(n-1)sd(e_{X|Z})^2}}} \\
&= \frac{\rho_{X,Y|Z}sd(e_{Y|Z})\sqrt{(n-p)(n-1)}}{\sqrt{\sum(e_{Y|Z} - \hat{e}_{Y|Z})^2}} \\
&= \frac{\rho_{X,Y|Z}\sqrt{(n-p)}\sqrt{(n-1)var(e_{Y|Z})}}{\sqrt{\sum(e_{Y|Z} - \hat{e}_{Y|Z})^2}} \\
&= \sqrt{(n-p)}\rho_{X,Y|Z}\sqrt{\frac{\sum(e_{Y|Z} - \bar{e}_{Y|Z})^2}{\sum(e_{Y|Z} - \hat{e}_{Y|Z})^2}}
\end{aligned}$$

Here we note that for a simple linear regression, $1-r^2 = \dfrac{SSE}{SST} = \dfrac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y}_i)^2}$, where r is the corresponding correlation coefficient. So the final term in the expression above simplifies to $\dfrac{1}{\sqrt{1-\rho^2_{X,Y|Z}}}$. And we have:

$$
\begin{aligned}
t_X &= \sqrt{(n-p)}\rho_{X,Y|Z}\sqrt{\frac{\sum(e_{Y|Z}-\bar{e}_{Y|Z})^2}{\sum(e_{Y|Z}-\hat{e}_{Y|Z})^2}} \\
&= \frac{\sqrt{(n-p)}\rho_{X,Y|Z}}{\sqrt{1-\rho^2_{X,Y|Z}}} \\
t_X\sqrt{1-\rho^2_{X,Y|Z}} &= \sqrt{(n-p)}\rho_{X,Y|Z} \\
t_X^2(1-\rho^2_{X,Y|Z}) &= (n-p)\rho^2_{X,Y|Z} \\
t_X^2 &= (n-p)\rho^2_{X,Y|Z} + t_X^2\rho^2_{X,Y|Z} \\
\rho^2_{X,Y|Z} &= \frac{t_X^2}{t_X^2+(n-p)} \\
\rho_{X,Y|Z} &= \frac{t_X}{\sqrt{t_X^2+(n-p)}}
\end{aligned}
$$

We now exchange Y and X in the equation $Y = D\beta' + \epsilon$ by defining $F = \begin{bmatrix} Y & Z \end{bmatrix}$, $\Gamma = \begin{bmatrix} \gamma_1 & \gamma_0 & \gamma_2 & \gamma_3 \end{bmatrix}$ and the corresponding equation $X = F\Gamma' + \delta$, we are able to equate the corresponding t-statistic for Y, t_Y, by noting that $\rho_{X,Y|Z} = \rho_{Y,X|Z}$:

$$
\begin{aligned}
\frac{t_X^2}{t_X^2+(n-p)} &= \frac{t_Y^2}{t_Y^2+(n-p)} \\
t_X^2(t_Y^2+(n-p)) &= t_Y^2(t_X^2+(n-p)) \\
t_X^2t_Y^2 + t_X^2(n-p) &= t_Y^2t_X^2 + t_Y^2(n-p) \\
t_X^2 &= t_Y^2
\end{aligned}
$$

Since the correlation coefficient relationship $\rho_{X,Y|Z} = \rho_{Y,X|Z}$ implies that both have the same sign, $\frac{t_X}{\sqrt{t_X^2 + (n-p)}}$ must have the same sign as the corresponding expression for t_Y. The denominators in both equivalent expressions must be positive, so the numerator determines the sign of the expression. It follows that $t_X = t_Y$.

7.3 Appendix C

7.3.1 Data Sets GSE19615, GSE11121, GSE7390

The three datasets GSE19615, GSE11121, GSE7390 are available from the Gene Expression Omnibus [26]. We obtained the datasets using the MetaGX package in R (available at https://github.com/bhaibeka/MetaGx). Their key characteristics are summarized in Tables 7.1–7.3 below. In our analyses, we dropped the two patients in GSE7390 whose tumor grade was unknown.

7.3.2 MammaPrint Prediction

We used the genefu package in R [38] to make MammaPrint predictions using the gene expression data supplied with each dataset described in Section 1. We specifically used the `gene70` function, which takes as input the expression data matrix and gene annotations and provides as output both a continuous risk score and the dichotomized risk classification. We used the latter as the MammaPrint risk covariate in our covariate adjustment steps. For each dataset, we used the same covariate sets $W_{-ER}, W_C, W_G, W_{CG}$ for adjustment, as defined in section 2.3 of the main text.

Characteristic	Summary
n	115
Age (years)	53.89 (11.78)
Five-Year Recurrence	
Yes	60
No	55
Tumor Size (cm)	2.31 (1.21)
Grade	
1	23
2	28
3	64
Unknown	0
ER	
+	70
−	45
Unknown	0
MammaPrint Risk Prediction	
High	87
Low	28

Table 7.1: Characteristics of dataset GSE19615. ER - estrogen receptor status, Grade - tumor severity grading (3 is most severe), Five-Year Recurrence - whether or not cancer has reappeared after five years, MammaPrint risk prediction - high or low risk for cancer recurrence. Age and Tumor Size are given as means with standard deviations in parentheses.

Characteristic	Summary
n	200
Age (years)	59.98 (12.36)
Five-Year Recurrence	
Yes	153
No	47
Tumor Size (cm)	2.07 (0.99)
Grade	
1	29
2	136
3	35
Unknown	0
ER	
+	162
−	38
Unknown	0
MammaPrint Risk Prediction	
High	142
Low	58

Table 7.2: Characteristics of dataset GSE11121. ER - estrogen receptor status, Grade - tumor severity grading (3 is most severe), Five-Year Recurrence - whether or not cancer has reappeared after five years, MammaPrint risk prediction - high or low risk for cancer recurrence. Age and Tumor Size are given as means with standard deviations in parentheses.

Characteristic	Summary
n	198
Age (years)	46.39 (7.22)
Five-Year Recurrence	
Yes	135
No	63
Tumor Size (cm)	2.18 (0.80)
Grade	
1	30
2	83
3	83
Unknown	2
ER	
+	134
−	64
Unknown	0
MammaPrint Risk Prediction	
High	144
Low	54

Table 7.3: Characteristics of dataset GSE7390. ER - estrogen receptor status, Grade - tumor severity grading (3 is most severe), Five-Year Recurrence - whether or not cancer has reappeared after five years, MammaPrint risk prediction - high or low risk for cancer recurrence. Age and Tumor Size are given as means with standard deviations in parentheses.

7.3.3 Differences between unadjusted and adjusted estimators

To assess how different the estimators computed under the unadjusted and adjusted cases are, we looked at the difference $\hat{\psi}^j_{una} - \hat{\psi}^j_{adj}$ over the $j = 1, \ldots, 100,000$ iterations in each of the four simulations using the four datasets in our study. A histogram of the differences for the simulation using the MammaPrint validation dataset is presented in the main manuscript. Three histograms for the simulations using GSE19615, GSE11121, and GSE7390 appear in this supplement, below. We also present in **Table 7.4** a comparison across all four studies of the average difference, the standard deviation of the difference, and the percentage of times that the unadjusted estimator was larger in absolute value than the adjusted estimator. Since the true treatment effect was set to zero in each simulation study, if the adjustment covariates are prognostic of the outcome, we would expect the adjusted estimator to be closer to zero more often than the unadjusted estimator. This occurred in over 50% of the iterations in all four studies. In all cases, we used the estimators adjusted for all available covariates (clinical + genomic).

7.3.4 Variation in magnitude of precision loss when covariates are not prognostic

We presented in **Table 4.3** of the main text the loss in precision due to adjustment when data was generated from a distribution with W and Y independent. We used more covariates than are usually recommended for this procedure because we wanted to include all clinically relevant baseline covariates that are usually measured for a breast cancer patient. We found that the sample size

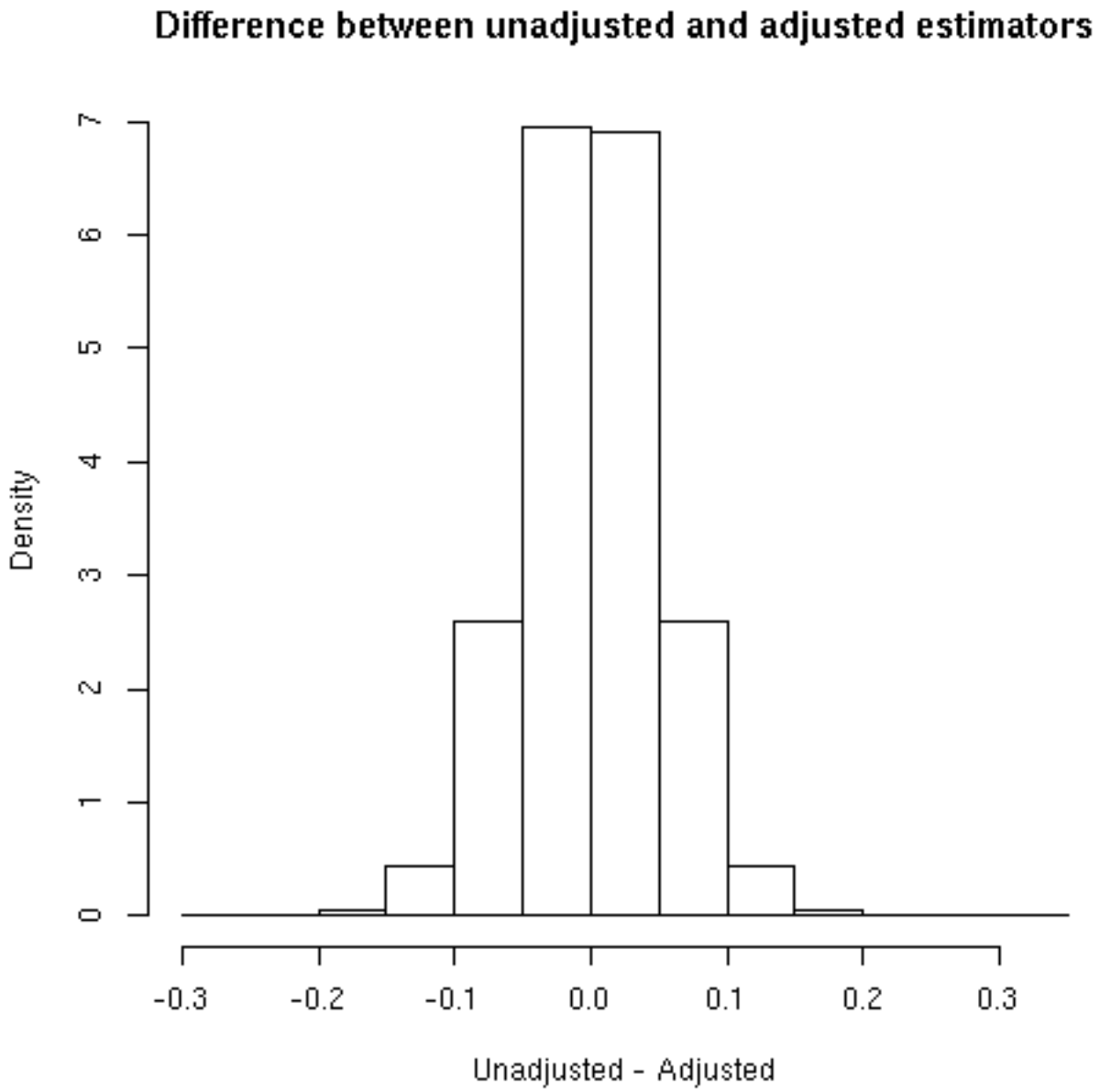

Figure 7.2: Histogram of $\hat{\psi}^j_{una} - \hat{\psi}^j_{adj}$, GSE19615. The histogram of differences between the unadjusted and adjusted estimators is roughly normal and is centered close to zero (mean=-6.7e-07, standard deviation=0.05). The unadjusted estimator is larger in absolute value than the adjusted estimator in approximately 55% of simulations.

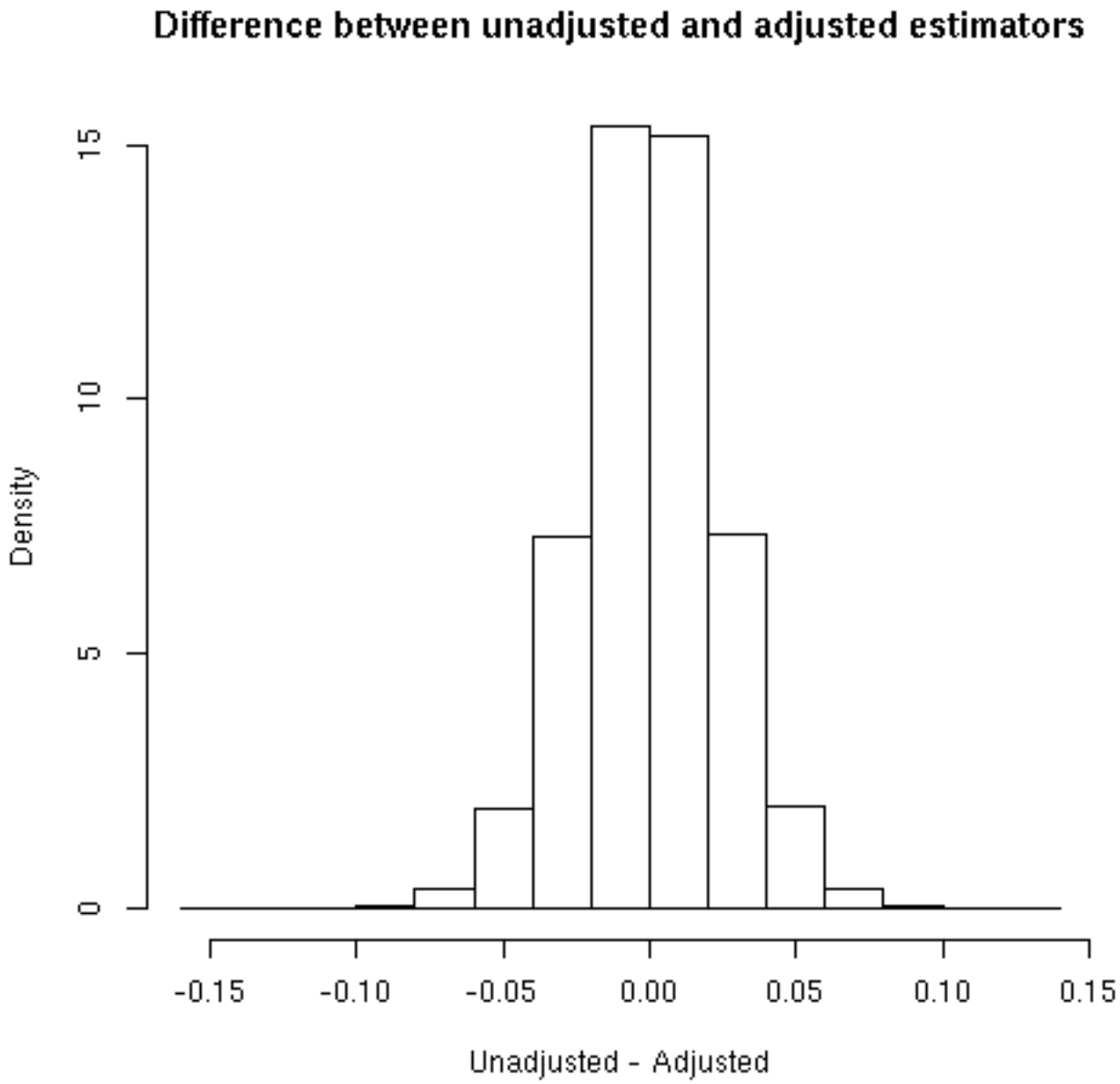

Figure 7.3: Histogram of $\hat{\psi}^j_{una} - \hat{\psi}^j_{adj}$, GSE11121. The histogram of differences between the unadjusted and adjusted estimators is roughly normal and is centered close to zero (mean=-4.6e-05, standard deviation=0.0242). The unadjusted estimator is larger in absolute value than the adjusted estimator in approximately 53% of simulations.

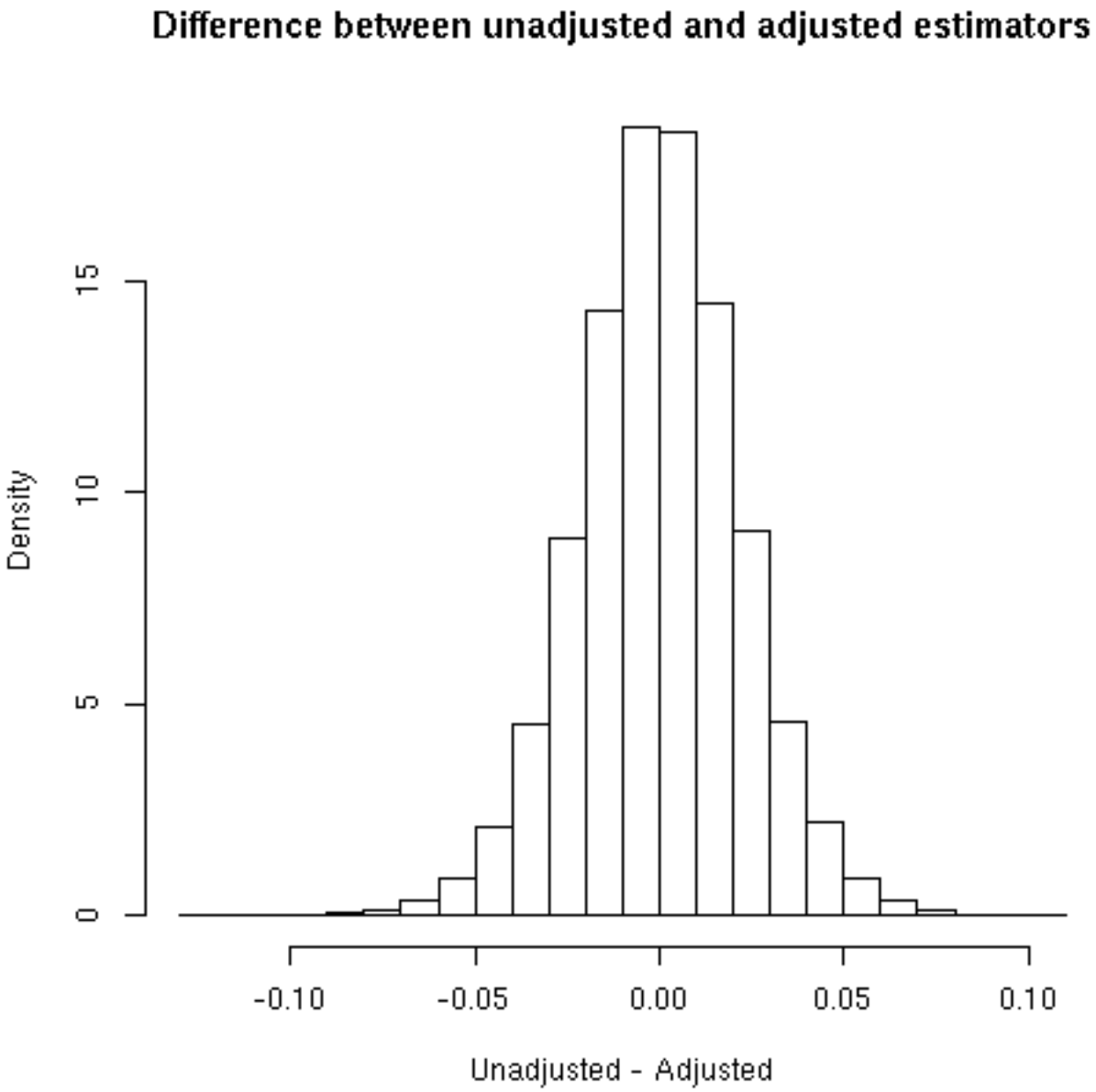

Figure 7.4: Histogram of $\hat{\psi}^j_{una} - \hat{\psi}^j_{adj}$, GSE7390. The histogram of differences between the unadjusted and adjusted estimators is roughly normal and is centered close to zero (mean=0.0001, standard deviation=0.0219). The unadjusted estimator is larger in absolute value than the adjusted estimator in approximately 51% of simulations.

Dataset	mean($\hat{\psi}^j_{una} - \hat{\psi}^j_{adj}$)	SD($\hat{\psi}^j_{una} - \hat{\psi}^j_{adj}$)	% $\lvert\hat{\psi}^j_{una}\rvert > \lvert\hat{\psi}^j_{adj}\rvert$
Mammaprint	5.4e-05	0.0145	53.1
GSE19615	-6.7e-07	0.05	55.2
GSE11121	-4.6e-05	0.0242	52.8
GSE7390	1.3e-04	0.0219	51.0

Table 7.4: Differences between unadjusted and adjusted estimators We find that the average difference between the unadjusted and adjusted estimators is similar across all simulations and the standard deviations are comparable, although the standard deviation in GSE19615 is more than twice as large as the others. The final column in the table shows the percentage of simulation iterations in which the adjusted estimator was closer in absolute value than the unadjusted estimator to the true treatment effect of zero. For each dataset, this occurred in slightly more than 50% of the iterations.

	Original Sample Size		
Covariate Set	σ^2_{una}	σ^2_{adj}	G_{adj}
W_{-ER}	0.00178	0.00180	-1.1%
W_C	0.00178	0.00181	-1.5%
W_G	0.00178	0.00179	-0.4%
W_{CG}	0.00178	0.00181	-1.8%

Table 7.5: Precision gains under data generating distribution with W and Y independent, based on marginal distributions from Mammaprint validation data set, using fewer clinical covariates.

in the simulated trials and the number of covariates we included affected the magnitude of precision losses. To illustrate, we conducted additional simulation studies where W and Y are independent, both using the MammaPrint validation dataset, where we used fewer adjustment covariates as shown in **Table 7.5**. Specifically, we defined new covariate sets W'_{-ER} ={Tumor Size}, W'_C = {Tumor Size, ER status}, W'_G = {MammaPrint Risk Prediction}, W'_{CG} = {Tumor Size, ER status, MammaPrint Risk Prediction}. The precision losses were smaller in magnitude when we reduced the number of adjustment covariates in this way.

Bibliography

[1] D. B. Allison, X. Cui, G. P. Page, and M. Sabripour. Microarray data analysis: from disarray to consolidation and consensus. *Nat. Rev. Genet.*, 7(1):55–65, Jan 2006.

[2] Samantha F Anderson and Scott E Maxwell. Theres more than one way to conduct a replication study: Beyond statistical significance. *Psychological Methods*, http://dx.doi.org/10.1037/met0000051, 2015.

[3] Ramy Arnaout, Thomas P Buck, Paulvalery Roulette, and Vikas P Sukhatme. Predicting the cost and pace of pharmacogenomic advances: an evidence-based study. *Clinical chemistry*, 59(4):649–657, 2013.

[4] Jens B Asendorpf, Mark Conner, Filip De Fruyt, Jan De Houwer, Jaap JA Denissen, Klaus Fiedler, Susann Fiedler, David C Funder, Reinhold Kliegl, Brian A Nosek, et al. Recommendations for increasing replicability in psychology. *European Journal of Personality*, 27(2):108–119, 2013.

[5] K. A. Baggerly, J. S. Morris, S. R. Edmonson, and K. R. Coombes. Signal in noise: evaluating reported reproducibility of serum proteomic tests for ovarian cancer. *J. Natl. Cancer Inst.*, 97(4):307–309, Feb 2005.

[6] Keith A Baggerly and Kevin R Coombes. Deriving chemosensitivity from cell lines: Forensic bioinformatics and reproducible research in high-throughput biology. *The Annals of Applied Statistics*, pages 1309–1334, 2009.

[7] Monya Baker. Over half of psychology studies fail reproducibility test, August 2015. http://www.nature.com/news/over-half-of-psychology-studies-fail-reproducibility-test-1.18248.

[8] AD Barker, CC Sigman, GJ Kelloff, NM Hylton, DA Berry, and LJ Esserman. I-SPY 2: An Adaptive Breast Cancer Trial Design in the Setting of Neoadjuvant Chemotherapy. *Clinical Pharmacology & Therapeutics*, 86(1):97–100, 2009.

[9] T. Barrett, S. E. Wilhite, P. Ledoux, C. Evangelista, I. F. Kim, M. Tomashevsky, K. A. Marshall, K. H. Phillippy, P. M. Sherman, M. Holko, A. Yefanov, H. Lee, N. Zhang, C. L. Robertson, N. Serova, S. Davis, and A. Soboleva. NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Res.*, 41(Database issue):D991–995, Jan 2013.

[10] Ben Baumer, Mine Cetinkaya-Rundel, Andrew Bray, Linda Loi, and Nicholas J Horton. R markdown: Integrating a reproducible analysis tool into introductory statistics. *arXiv preprint arXiv:1402.1894*, 2014.

[11] H. Bengtsson, K. Simpson, J. Bullard, and K. Hansen. aroma.affymetrix: A generic framework in R for analyzing small to very large Affymetrix data sets in bounded memory. Technical Report 745, Department of Statistics, University of California, Berkeley, February 2008.

[12] Donald A Berry. Adaptive clinical trials: the promise and the caution. *Journal of Clinical Oncology*, 29(6):606–609, 2011.

[13] Bernd Bischl, Michel Lang, Olaf Mersmann, Joerg Rahnenfuehrer, and Claus Weihs. Computing on high performance clusters with R: Packages BatchJobs and BatchExperiments. Technical Report 1, TU Dortmund, 2011.

[14] Yvonne Bombard, Peter B Bach, and Kenneth Offit. Translating genomics in cancer care. *Journal of the National Comprehensive Cancer Network*, 11(11):1343–1353, 2013.

[15] Sanford L Braver, Felix J Thoemmes, and Robert Rosenthal. Continuously cumulating meta-analysis and replicability. *Perspectives on Psychological Science*, 9(3):333–342, 2014.

[16] Wylie Burke, Diane M Korngiebel, and Mildred Cho. Closing the gap between knowledge and clinical application: Challenges for genomic translation. *PLoS genetics*, 11(2):e1004978–e1004978, 2015.

[17] Marc Buyse, Sherene M Loi, Laura Van't Veer, Giuseppe Viale, Mauro Delorenzi, Annuska M. Glas, Mahasti Saghatchian d'Assignies, Jonas Bergh, Rosette Lidereau, Paul Ellis, Adrian Harris, Jan Bogaerts, Patrick Therasse, Arno Floore, Mohamed Amakrane, Fanny Piette, Emiel Rutgers, Christos Sotiriou, Fatima Cardoso, and Martine J Piccart. Validation and Clinical Utility of a 70-Gene Prognostic Signature for Women With Node-Negative Breast Cancer. *Journal of the National Cancer Institute*, 98(17):1183–1192, 2006.

[18] W. Cao, A.A. Tsiatis, and M. Davidian. Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. *Biometrika*, 96:723–34, 2009.

[19] Lon R Cardon and John I Bell. Association study designs for complex diseases. *Nature Reviews Genetics*, 2(2):91–99, 2001.

[20] Mary Cianfrocca and Lori J. Goldstein. Prognostic and predictive factors in early-stage breast cancer. *The Oncologist*, 9(6):606–616, 2004.

[21] Elizabeth Colantuoni and Michael Rosenblum. Leveraging prognostic baseline variables to gain precision in randomized trials. *Statistics in Medicine*, 34(18):2602–2617, 2015.

[22] Open Science Collaboration et al. Estimating the reproducibility of psychological science. *Science*, 349(6251):aac4716, 2015.

[23] Francis S Collins and Harold Varmus. A new initiative on precision medicine. *New England Journal of Medicine*, 372(9):793–795, 2015.

[24] R. M. Connolly and A. C. Wolff. Omics as useful tools in clinical practice: are we there yet? *Oncology (Williston Park, N.Y.)*, 27(3):216–218, Mar 2013.

[25] Roisin M Connolly. Omics as useful tools in clinical practice: are we there yet? *ONCOLOGY*, 27(3), 2013.

[26] Ron Edgar, Michael Domrachev, and Alex E Lash. Gene expression omnibus: Ncbi gene expression and hybridization array data repository. *Nucleic acids research*, 30(1):207–210, 2002.

[27] Ronald A Fisher. Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika*, pages 507–521, 1915.

[28] Deborah Ford, Douglas F Easton, D Timothy Bishop, Steven A Narod, and David E Goldgar. Risks of cancer in brca1-mutation carriers. *The Lancet*, 343(8899):692–695, 1994.

[29] Andrew Gelman and John Carlin. Beyond power calculations assessing type s (sign) and type m (magnitude) errors. *Perspectives on Psychological Science*, 9(6):641–651, 2014.

[30] Andrew Gelman and Eric Loken. The statistical crisis in science. *American Scientist*, 102(6):460, 2014.

[31] Andrew Gelman and David Weakliem. Of beauty, sex and power: Too little attention has been paid to the statistical challenges in estimating small effects. *American Scientist*, pages 310–316, 2009.

[32] D. Geman, C. d'Avignon, D. Q. Naiman, and R. L. Winslow. Classifying gene expression profiles from pairwise mRNA comparisons. *Stat Appl Genet Mol Biol*, 3:Article19, 2004.

[33] A. M. Glas, A. Floore, L. J. Delahaye, A. T. Witteveen, R. C. Pover, N. Bakx, J. S. Lahti-Domenici, T. J. Bruinsma, M. O. Warmoes, R. Bernards, L. F. Wessels, and L. J. Van't Veer. Converting a breast cancer microarray signature into a high-throughput diagnostic test. *BMC Genomics*, 7:278, 2006.

[34] I Gnarra et al. Mutations of the vhl tumour suppressor gene in renal. *Nat Genet*, 7:85–90, 1994.

[35] Steven N Goodman. A comment on replication, p-values and evidence. *Statistics in medicine*, 11(7):875–879, 1992.

[36] S. Gruber and M.J. van der Laan. Targeted minimum loss based estimator that outperforms a given estimator. *The International Journal of Biostatistics*, 8(1):Article 11, 2012.

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

[38] B Haibe-Kains, M Schroeder, G Bontempi, C Sotiriou, and J Quackenbush. genefu: relevant functions for gene expression analysis, especially in breast cancer. r package version 191, 2012.

[39] Benjamin Haibe-Kains, Markus Schroeder, Aedin C. Culhane, Gianluca Bontempi, Christos Sotiriou, and John F. Quackenbush. GeneFu: Relevant Functions for Gene Expression Analysis, Especially in Breast Cancer. Technical report, 2011.

[40] T. Hastie, R. Tibshirani, Balasubramanian Narasimhan, and Gil Chu. *pamr: Pam: prediction analysis for microarrays*, 2014. R package version 1.55.

[41] Jim Hester. *knitrBootstrap: Knitr Bootstrap framework.*, 2014. R package version 1.0.0.

[42] John PA Ioannidis. Contradicted and initially stronger effects in highly cited clinical research. *Jama*, 294(2):218–228, 2005.

[43] Leah R Jager and Jeffrey T Leek. An estimate of the science-wise false discovery rate and application to the top medical literature. *Biostatistics*, 15(1):1–12, 2014.

[44] C. Jennison and B.W. Turnbull. *Group Sequential Methods with Applications to Clinical Trials*. Chapman and Hall/CRC Press, Boca Raton, FL, 1999.

[45] Alison Ledgerwood. Introduction to the special section on advancing our methods and practices. *Perspectives on Psychological Science*, 9(3):275–277, 2014.

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

[47] Jeffrey T Leek, W Evan Johnson, Hilary S Parker, Andrew E Jaffe, and John D Storey. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics*, 28(6):882–883, 2012.

[48] Jeffrey T Leek and Roger D Peng. Statistics: P values are just the tip of the iceberg. *Nature*, 520(7549):612–612, 2015.

[49] Benjamin Letham, Cynthia Rudin, Tyler H McCormick, and David Madigan. Interpretable classifiers using rules and Bayesian analysis: Building a better stroke prediction model. *Annals of Applied Statistics*, In Press.

[50] The Cancer Letter. Duke Accepts Potti Resignation; Retraction Process Initiated with Nature Medicine, November 2011.

[51] Danny Liaw, Debbie J Marsh, Jing Li, Patricia LM Dahia, Steven I Wang, Zimu Zheng, Shikha Bose, Katherine M Call, Hui C Tsou, Monica Peacoke, et al. Germline mutations of the pten gene in cowden disease, an inherited breast and thyroid cancer syndrome. *Nature genetics*, 16(1):64–67, 1997.

[52] Lara Lusa, Lisa M. McShane, James F Reid, Loris De Cecco, Federico Ambrogi, Elia Biganzoli, Manuela Gariboldi, and Marco A Pierotti. Challenges in Projecting Clustering Results Across Gene Expression Profiling Datasets. *Journal of the National Cancer Institute*, 99(22):1715–1723, 2007.

[53] I. J. Majewski and R. Bernards. Taming the dragon: genomic biomarkers to individualize the treatment of cancer. *Nat. Med.*, 17(3):304–312, Mar 2011.

[54] L. Marchionni, B. Afsari, D. Geman, and J. T. Leek. A simple and reproducible breast cancer prognostic test. *BMC Genomics*, 14:336, 2013.

[55] Luigi Marchionni, Bahman Afsari, Donald Geman, and Jeffrey T Leek. A simple and reproducible breast cancer prognostic test. *BMC genomics*, 14(1):336, 2013.

[56] Scott E Maxwell. The persistence of underpowered studies in psychological research: causes, consequences, and remedies. *Psychological methods*, 9(2):147, 2004.

[57] Matthew N McCall, Benjamin M Bolstad, and Rafael A Irizarry. Frozen robust multiarray analysis (frma). *Biostatistics*, 11(2):242–253, 2010.

[58] Blakeley B McShane and Ulf Böckenholt. You cannot step into the same river twice when power analyses are optimistic. *Perspectives on Psychological Science*, 9(6):612–625, 2014.

[59] S. Michiels, S. Koscielny, and C. Hill. Prediction of cancer outcome with microarrays: a multiple random validation strategy. *Lancet*, 365(9458):488–492, 2005.

[60] John Muschelli. *diffr: Display Differences Between Two Files using Codediff Library*, 2015. R package version 0.0.1.

[61] Soonmyung Paik, Steven Shak, Gong Tang, Chungyeul Kim, Joffre Baker, Maureen Cronin, Frederick L Baehner, Michael G Walker, Drew Watson, Taesung Park, et al. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *New England Journal of Medicine*, 351(27):2817–2826, 2004.

[62] Hilary S Parker, Héctor Corrada Bravo, and Jeffrey T Leek. Removing batch effects for prediction problems with frozen surrogate variable analysis. *arXiv preprint arXiv:1301.3947*, 2013.

[63] J. S. Parker, M. Mullins, M. C. Cheang, S. Leung, D. Voduc, T. Vickery, S. Davies, C. Fauron, X. He, Z. Hu, J. F. Quackenbush, I. J. Stijleman, J. Palazzo, J. S. Marron, A. B. Nobel, E. Mardis, T. O. Nielsen, M. J. Ellis, C. M. Perou, and P. S. Bernard. Supervised risk predictor of breast cancer based on intrinsic subtypes. *J. Clin. Oncol.*, 27(8):1160–1167, Mar 2009.

[64] H Parkinson, M Kapushesky, M Shojatalab, N Abeygunawardena, R Coulson, A Farne, E Holloway, N Kolesnykov, P Lilja, M Lukk, R Mani, T Rayner, A Sharma, E William, U Sarkans, and A Brazma. ArrayExpress–a public database of microarray experiments and gene expression profiles. *Nucleic acids research*, 35(Database issue):D747–50, January 2007.

[65] Prasad Patil, Pierre-Olivier Bachant-Winner, Benjamin Haibe-Kains, and Jeffrey T Leek. Test set bias affects reproducibility of gene signatures. *Bioinformatics*, page btv157, 2015.

[66] Prasad Patil and Jeffrey T. Leek. Reporting of 36% of studies replicate in the media, September 2015. https://github.com/jtleek/replication_paper/blob/gh-pages/in_the_media.md.

[67] Roger D Peng. Reproducible research in computational science. *Science (New York, Ny)*, 334(6060):1226, 2011.

[68] Roger D Peng, Francesca Dominici, and Scott L Zeger. Reproducible epidemiologic research. *American journal of epidemiology*, 163(9):783–789, 2006.

[69] E. F. Petricoin, A. M. Ardekani, B. A. Hitt, P. J. Levine, V. A. Fusaro, S. M. Steinberg, G. B. Mills, C. Simone, D. A. Fishman, E. C. Kohn, and L. A. Liotta. Use of proteomic patterns in serum to identify ovarian cancer. *Lancet*, 359(9306):572–577, Feb 2002.

[70] Stephen R Piccolo, Ying Sun, Joshua D Campbell, Marc E Lenburg, Andrea H Bild, and W Evan Johnson. A single-sample microarray normalization method to facilitate personalized-medicine workflows. *Genomics*, 100(6):337–344, 2012.

[71] Xavier Robin, Natacha Turck, Alexandre Hainard, Natalia Tiberti, Frédérique Lisacek, Jean-Charles Sanchez, and Markus Müller. proc: an open-source package for r and s+ to analyze and compare roc curves. *BMC bioinformatics*, 12(1):1, 2011.

[72] J.M. Robins, M. Sued, Q. Lei-Gomez, and A. Rotnizky. Comment: Performance of double-robust estimators when inverse probability weights are highly variable. *Statistical Science*, 22(4):544–59, 2007.

[73] Andrea Rotnitzky, Quanhong Lei, Mariela Sued, and James M Robins. Improved double-robust estimation in missing data and causal inference models. *Biometrika*, 99(2):439–456, 2012.

[74] Felix D Schönbrodt and Marco Perugini. At what sample size do correlations stabilize? *Journal of Research in Personality*, 47(5):609–612, 2013.

[75] P. Sebastiani, N. Solovieff, A. Puca, S. W. Hartley, E. Melista, S. Andersen, D. A. Dworkis, J. B. Wilk, R. H. Myers, M. H. Steinberg, M. Montano, C. T. Baldwin, and T. T. Perls. Genetic signatures of exceptional longevity in humans. *Science*, 2010, Jul 2010.

[76] Daniel J Simons, Alex O Holcombe, and Barbara A Spellman. An introduction to registered replication reports at perspectives on psychological science. *Perspectives on Psychological Science*, 9(5):552–555, 2014.

[77] Peter H St George-Hyslop, Rudolph E Tanzi, Ronald J Polinsky, Jonathan L Haines, Linda Nee, Paul C Watkins, Richard H Myers, Robert G Feldman, Daniel Pollen, David Drachman, et al. The genetic defect causing familial alzheimer's disease maps on chromosome 21. *Science*, 235(4791):885–890, 1987.

[78] David J Stanley and Jeffrey R Spence. Expectations for replications are yours realistic? *Perspectives on Psychological Science*, 9(3):305–318, 2014.

[79] A. C. Tan, D. Q. Naiman, L. Xu, R. L. Winslow, and D. Geman. Simple decision rules for classifying human cancers from gene expression profiles. *Bioinformatics*, 21(20):3896–3904, Oct 2005.

[80] Z. Tan. Bounded, efficient and doubly robust estimating equations for marginal and nested structural models. *Biometrika*, 97:661–82, 2010.

[81] Terry Therneau, Beth Atkinson, and Brian Ripley. *rpart: Recursive Partitioning*, 2013. R package version 4.1-4.

[82] Robert Tibshirani, Trevor Hastie, Balasubramanian Narasimhan, and Gilbert Chu. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences*, 99(10):6567–6572, 2002.

[83] Marc J Van De Vijver, Yudong D He, Laura J van't Veer, Hongyue Dai, Augustinus AM Hart, Dorien W Voskuil, George J Schreiber, Johannes L Peterse, Chris Roberts, Matthew J Marton, et al. A gene-expression signature as a predictor of survival in breast cancer. *New England Journal of Medicine*, 347(25):1999–2009, 2002.

[84] L. J. van 't Veer, H. Dai, M. J. van de Vijver, Y. D. He, A. A. Hart, M. Mao, H. L. Peterse, K. van der Kooy, M. J. Marton, A. T. Witteveen, G. J. Schreiber, R. M. Kerkhoven, C. Roberts, P. S. Linsley, R. Bernards, and S. H. Friend. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415(6871):530–536, Jan 2002.

[85] Peter M Visscher, Matthew A Brown, Mark I McCarthy, and Jian Yang. Five years of gwas discovery. *The American Journal of Human Genetics*, 90(1):7–24, 2012.

[86] Levi Waldron, Benjamin Haibe-Kains, Aedín C Culhane, Markus Riester, Jie Ding, Xin Victoria Wang, Mahnaz Ahmadifar, Svitlana Tyekucheva, Christoph Bernau, Thomas Risch, et al. Comparative meta-analysis of prognostic gene signatures for late-stage ovarian cancer. *Journal of the National Cancer Institute*, 106(5):dju049, 2014.

[87] Jacob Westfall, David A Kenny, and Charles M Judd. Statistical power and optimal design in experiments in which samples of participants respond to samples of stimuli. *Journal of Experimental Psychology: General*, 143(5):2020, 2014.

[88] Richard Wooster, Graham Bignell, Jonathan Lancaster, Sally Swift, Sheila Seal, Jonathan Mangion, Nadine Collins, Simon Gregory, Curtis Gumbs, Gos Micklem, et al. Identification of the breast cancer susceptibility gene brca2. *Nature*, 378(6559):789–792, 1995.

[89] Elizabeth A Worthey, Alan N Mayer, Grant D Syverson, Daniel Helbling, Benedetta B Bonacci, Brennan Decker, Jaime M Serpe, Trivikram Dasu, Michael R Tschannen, Regan L Veith, et al. Making a definitive diagnosis: successful clinical application of whole exome sequencing in a child with intractable inflammatory bowel disease. *Genetics in Medicine*, 13(3):255–262, 2011.

[90] Yihui Xie. knitr: A general-purpose package for dynamic report generation in r. *R package version*, 1(7), 2013.

[91] L. Yang and A.A. Tsiatis. Effciency study of estimators for a treatment effect in a pretest-posttest trial. *The American Statistician*, 55:314–321, 2001.

Prasad Patil

1201 W. Mount Royal Avenue, Apartment 523
Baltimore, MD 21217
(978) 760-2736
prpatil42@gmail.com

EDUCATION

Ph.D. in Biostatistics
Johns Hopkins University, Baltimore, MD, 2011-2016
Advisor: Jeffrey Leek
Dissertation Title: "Assessing Reproducibility and Value in Genomic Signatures"

B.A. in Mathematics
New York University, New York, NY, 2005-2008
Concentration: Applied Mathematics
Minor: Computer Science

RESEARCH EXPERIENCE

Graduate Research 2012-2016
Department of Biostatistics, Johns Hopkins University, Baltimore, MD

With Jeff Leek,

- Identified and described bias due to data normalization in the prediction of breast cancer subtypes using gene expression information.
- Formalized feature selection and modeling with Top-Scoring Pairs for simple, decision-tree-based classifiers.
- Developed R packages for standardized analysis templating (`tdsm`) and interactive health visualizations (`healthvis`).
- Suggested 95% prediction intervals as a means of assessing whether a study result has been replicated.

With Jeff Leek and Michael Rosenblum,

- Used RCT baseline covariate adjustment methods to assess the additional value a genomic prediction can provide beyond standard clinical measurements in improving the precision of a treatment effect estimator.
- Examined the benefit of using machine learning methods to summarize the predictive value of large numbers of baseline covariates in an RCT setting.

Scientific Programmer 2009-2011
Center for Biomedical Informatics, Harvard Medical School, Boston, MA

- Developed synthetic patient simulation platform predicated on Bayesian networks, and accompanying web service.
- Built clinical trial simulation framework using synthetic patients, stochastic PK-PD models for drug clearence, and coded clinical trial protocols.
- Fosterd collaboration with Beth Israel Deaconess Medical Center and GenomeQuest, Inc. to create a commercial, clinical-grade omics analysis pipeline intended for hospital use.
- Administered a cloud computing environment via Amazon Web Services for day-to-day lab operations and omics pipeline deployement.

Bioinformatics Fellow 2008-2009
Partners Center for Personalized Genetic Medicine, Boston, MA

- Collaborated with the Center for Biomedical Informatics at Harvard Medical School to develop an open-source pipeline using next-generation sequencing (NGS) technology to detect and clincally annotate all variants in an individual human genome.
- Evaluated and tested NGS tools and restructured code for parallelization via Hadoop/MapReduce.
- Developed a comprehensive data model and modular scripted pipeline for variant annotation and reporting.

PUBLICATIONS

1. **Patil P**, Peng RD, Leek JT (2015). A glass half-full interpretation of replicability in psychological science. Revised and resubmitted, *Perspectives on Psychological Science*.
2. **Patil P**, Colantuoni E, Rosenblum MA, Leek JT (2015). Genomic and clinical predictors for improving estimator precision in randomized trials of breast cancer treatments. Accepted, *Contemporary Clinical Trials Communications*.
3. **Patil P**, Leek JT (2015). Discussion of "Visualizing statistical models: Removing the blindfold". *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 8(4), 240-241.
4. **Patil P**, Bachant-Winner PO, Haibe-Kains B, Leek JT (2015). Test set bias affects reproducibility of gene signatures. *Bioinformatics*, btv157.
5. Hyland PL, Burke LS, Pfeiffer RM, Rotunno M, Sun D, **Patil P**, Wu X, Tucker MA, Goldstein AM, Yang XR (2014). Constitutional promoter methylation and risk of familial melanoma. *Epigenetics*, 9(5), 685-692.
6. Fusaro, VA, **Patil P**, Chi CL, Contant CF, Tonellato PJ (2013). A Systems Approach to Designing Effective Clinical Trials Using Simulations. *Circulation*, 127(4), 517-526.
7. Fusaro VA, **Patil P**, Gafni E, Wall DP, Tonellato PJ (2011). Biomedical Cloud Computing Using Amazon Web Services. *PLoS Computational Bioogy.* 7(8):e1002147.
8. Wall DP, Kudtarkar P, Fusaro VA, Pivovarov R, **Patil P**, Tonellato PJ (2010). Cloud computing for comparative genomics. *BMC Bioinformatics* 11:259.

PRESENTATIONS AND POSTERS

1. **Patil P***, Alquicira J, Leek JT. Measuring the value of GWAS results in a clinical trial setting. ASHG 2015 [poster].
2. **Patil P***. What to expect when you're replicating. Hopkins Biostatistics Journal Club 2015 [talk].
3. **Patil P***, Leek JT. Reproducibility and value of genomic signatures. JSM 2015 [talk].
4. **Patil P***. On organization. Hopkins Biostatistics Computing Club 2015 [talk].
5. **Patil P***, Leek JT. Assessing the reproducibility and value of genomic signatures. ENAR 2015 [talk].
6. **Patil P***, Haibe-Kains B, Leek JT. Cross-platform gene signature development using Top-Scoring Pairs. JSM 2014 [talk].
7. **Patil P***. Cross-validation in the presence of many features. Hopkins Biostatistics Journal Club 2014 [talk].

8. **Patil P***, Haibe-Kains B, Leek JT. Cross-platform gene signature development using Top-Scoring Pairs. ENAR 2014 [talk].
9. Chi CL, Fusaro VA, **Patil P**, Crawford MA, Contant CF, Tonellato PJ. An approach to optimal individualized warfarin treatment through clinical trial simulations. Proceedings of IEEE Cairo International Biomedical Engineering Conference (CIBEC) 2010, Cairo, Egypt [talk].
10. Chi, CL, **Patil P**, Fusaro VA, Kos PJ, Pivovarov R, Contant CF, Tonellato PJ. A Simulation Platform to Examine Heterogeneity Influence on Treatment. Proceedings of the 2010 American Medical Informatics Association Annual Symposium, Washington, DC [talk].
11. Chi CL, Kos PJ, Fusaro VA, Pivovarov R, **Patil P**, Tonellato PJ. Mining personalized medicine algorithms with surrogate algorithm tags. Proceedings of the First ACM International Health Informatics Symposium 2010 [poster].
12. **Patil P***, Heus H, Arnaout R, Tonellato PJ. Refining a method for processing an individuals whole genome to clinical utility. CSHL Personal Genomes Meeting 2010. Cold Spring Harbor, N.Y [talk].
13. **Patil P***, Tonellato PJ. Individual Whole Genome Mapping: from NGS reads to clinical variants, American Medical Informatics Association Clinical Research Informatics Summit 2010, San Francisco, CA [poster].
14. Fusaro VA, Kos PJ, Tector M, Tector A, **Patil P**, Tonellato PJ. Electronic Medical Record Analysis Using Cloud Computing, American Medical Informatics Association Clinical Research Informatics Summit 2010, San Francisco, CA [poster].
15. **Patil P*.**Clinical algorithms for whole genome data Partners Health Care Information Systems Research Council Symposium 2009, Harvard Medical School, Boston, MA [talk].
16. **Patil P*.** Clinical annotation of an individual whole genome assembly, Center for Biomedical Informatics Research Day 2009, Harvard Medical School, Boston, MA [talk].

* - Presenter

TEACHING

Guest lecturer, BIO622 — 2015
Conducted daily lecture for 500+ student course

Lead TA, BIO621-623 — 2014-2015
Prepared and held 2-3 sections per week, 40-60 students each
Beta-tested, proctored, and graded exams

TA, BIO611-612, 615, 621-624, AS.280.35 — 2012-2014
Graded homework and exams, held small sections and office hours

SERVICE

Refereeing: Biometrics, Genome Biology, PLOS ONE

Organization

- Session chair, Next Generation Sequencing. ENAR 2014.
- Organizer (2012), Hopkins Biostatistics Computing Club.

AWARDS

Jane and Steve Dykacz Award 2016
Departmental award for best student paper in medcial statistics
Awarded for "Genomic and clinical predictors for improving estimator precision in randomized trials of breast cancer treatments"

Helen Abbey Award 2016
Departmental award for teaching

JHSPH Student Assembly Teaching Assistant Recognition Award 2015
One of two voted on by students across all courses in JHSPH

NYU Honors Scholar 2008

NYU Deans List 2006-2007

National Merit Scholarship 2005-2008

SOFTWARE

R Packages (submitted to Bioconductor):

`healthvis` (https://github.com/prpatil/healthvis)
Interactive health visualizations. Built using `d3`, `shiny`, `htmlWidgets`.

`tdsm` (https://github.com/prpatil/tdsm)
Templated Deterministic Statistical Machines. Automated analysis templates and standardized reports that can be edited and compared.

Languages:
R, Javascript, C/C++, Java, Perl, MATLAB, Stata, Hadoop, Shell scripting

REFERENCES

Available upon request.