Analyzing Biomarker Discovery: Estimating the Reproducibility of Biomarkers

by

Amirhosein Forouzandehmoghadam

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

Department of Computing Science

University of Alberta

Abstract

A *biomarker* is a feature (*e.g.*, gene expression, SNP, etc.) that is significantly different between two classes of instances – typically case and control. Knowing these biomarkers can help us understand a biological condition or identify the appropriate treatment for a certain disease. Many researchers try to identify these biomarkers by using univariate hypothesis testing over a labeled dataset – selecting a feature if it is statistically significantly different. However, such sets of proposed biomarkers are often not reproducible – subsequent studies typically fail to identify the same sets; indeed, there is often a very small overlap between the biomarkers proposed in various pairs of related studies, exploring the same phenotypes over the same distribution of subjects.

This dissertation first defines the *Reproducibility Score* for a labeled dataset, as a measure (in [0,1]) of reproducibility of the results produced by the specified biomarker discovery process, for this distribution of subjects. We then provide ways to reliably estimate this score – giving ways to produce an over-bound, an under-bound and a middle-value approximation for this score for a given dataset. These specific tools apply to the univariate hypothesis testing on dichotomous groups. We confirm that these approximations are meaningful by providing empirical results for many datasets (microarray, RNAseq and SNP), and show that these predictions match known reproducibility results. Finally, we explore how changing some of the settings of a biomarker discovery process (such as p-value threshold, p-value correction method, sample size, etc.) can affect the results and the *Reproducibility Score* using real datasets.

Preface

This dissertation is a collaborative research paper, "Analyzing Biomarker Discovery: Estimating the Reproducibility of Biomarkers" by Amir Forouzandeh, Alex Rutar, Sunil Kalmady and Russell Greiner.
The "Overbound" and "Underbound" algorithms are based on earlier work done by Dr. Sunil Kalmady and the "Middle-Value Approximation" is a continuation on the work done by Alex Rutar.

Acknowledgements

First and foremost I would like to thank my supervisor, Professor Russell Greiner, who has helped me through every step of this journey and for his endless patience and his guidance. This work would not have come to fruition without his support. He consistently allowed this dissertation to be my own work, but steered me in the right direction whenever he thought I needed it.

I would also like to thank Dr. Sunil Kalmady, who always offered his help whenever I needed, and made sure that I am headed in the right direction. His assistance and input were integral to overcoming many obstacles along the way.

I would like to thank all my friends in Edmonton for their support during this time, specially Mehran. I could not have done this without their help.

Finally, I must express my profound gratitude to my parents for providing me with unfailing support and encouragement throughout my years of study. This would not have been possible without them. Thank you.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Better understanding of a disease will clearly lead to better diagnosis and treatments. This often begins by finding "biomarkers", which generally refer to "a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention" [52]. These are typically individual features (*e.g.*, expression values of specific genes [54], [56]) that follow different distributions (*e.g.*, have different mean values) in diseased versus healthy subjects.

Biomedical researchers can sometimes identify these biomarkers based on their domain knowledge of the disease etiology and/or cellular pathways – seeking features that are causally related to the disease; perhaps corresponding to the cause of the disease (*e.g.*, phenylketonuria is caused by a single gene PAH, which codes for hepatic enzyme phenylalanine hydroxylase [7]) or a symptom of it (*e.g.*, Hemoglobin A1C for monitoring the degree of glucose metabolism in diabetes [31].) This dissertation, however, focuses on ways to evaluate and validate the biomarkers *discovered* from a given labeled dataset of earlier subjects – think of a matrix whose rows each correspond to a person, and whose columns each correspond to a feature (*e.g.*, clinical measure, or the expression value of a gene), with the final column being the label (*e.g.*, case versus control); see Figure 1.1.[1] These "biomarker discovery studies" (aka "association studies"[2]) then try to determine which of the features (columns) are statistically "different" in case versus control. This often involves first computing some statistics for each feature – *e.g.*, for real-valued entries, running a t-test based on the mean and variance over the controls and over the cases – then declaring a feature to be a biomarker if the resulting FDR-corrected

[1] For notation: We will refer to each of the first r columns of the matrix shown in Figure 1.1 as a "feature"; these are often called "(independent) variables". We will refer to the final column as a "label" – *e.g.*, case versus control – these are often called "dependent variables", "groups", "phenotypes" or "classes". Finally, we will use "instance" to refer to each row of that matrix; these are sometimes called "subjects" or "samples".

[2] Two standard examples here are the "Genome Wide Association Study" [GWAS], over a set of SNPs [10]; and the "Gene Signature Study" [8], among many others.

p-value is below 0.05 [62] – or stated more precisely, whenever we can reject the null hypothesis that the two means are equal; see terms defined below.

In some situations, the researchers then apply some biological or medical process to validate these biomarkers – *e.g.*, based on knock-out or amplification studies [15], [49]. Similarly, they may match the proposed biomarkers to existing biological knowledge – perhaps based on earlier knock-out studies. (Of course, this relies on having such prior knowledge, and knowing that it is correct.[3]) Alternatively, other projects instead use the purported biomarkers in a computational model – perhaps a classifier [1], [21], [60], then apply some measure on that down-stream model (such as its accuracy), and declare the biomarkers to be useful if that model scores well. A great many papers, however, simply publish the list of purported biomarkers, but provide no validation for this set.[4] This dissertation addresses this limitation by providing a falsifiable (statistical) claim about these biomarkers, which suggests a validation of these proposed biomarker sets.

While some biomarkers are causally related to the associated label, this is difficult to obtain (often requiring instrumented studies [40]), but fortunately, it may be sufficient for the features to be *correlated* with the phenotype. Here, an ideal biomarker discovery process would identify all-and-only the features that are *consistently* correlated with the associated disease, in that its presence (or absence or specified minimum concentration or ...) alone supports that disease. Hence, many researchers would say that a proposed biomarker is good if it is *reproducible* – *i.e.*, that the biomarkers found in one study, would appear in many (ideally, all) future studies that explore this disease. This has motivated the use of independent test sets to check the validity of the earlier findings. Unfortunately, many papers report this is not the case – *i.e.*, that relatively few biomarkers appear across multiple studies. For example, while the breast cancer studies by van't Veer *et al.* [54] (resp., Wang *et al.* [56]) reported signatures with 70 (resp., 76) genes, they had only 3 genes in common. Others have noticed this: Ein-Dor *et al.* [18] notes: "Only 17 genes appeared in both the list of 456 genes of Sorlie *et al.* [50] and the 231 genes of vant Veer *et al.* [54]; merely 2 genes were shared between the sets of Sorlie *et al.* and Ramaswamy *et al.*[42]. Such disparity is not limited to breast cancer but characterizes other human disease datasets (Lossos *et al.* [2]) such as schizophrenia (Miklos and Maleszka [35])". We should also note that observing certain associations in one dataset, does not mean that we would find the same associations in other independent datasets [36]. In fact, it has been shown that most promising biomarkers in one dataset will not have as good results in independent datasets [24], [45].

[3] And of course, studies that only replicate what is already known, will not identify novel biomarkers.

[4] This may be because biological validation is not yet implementable, or the technology is not yet available. Alternatively, the biological validation may be possible but this will be a major project to be explored in future works.

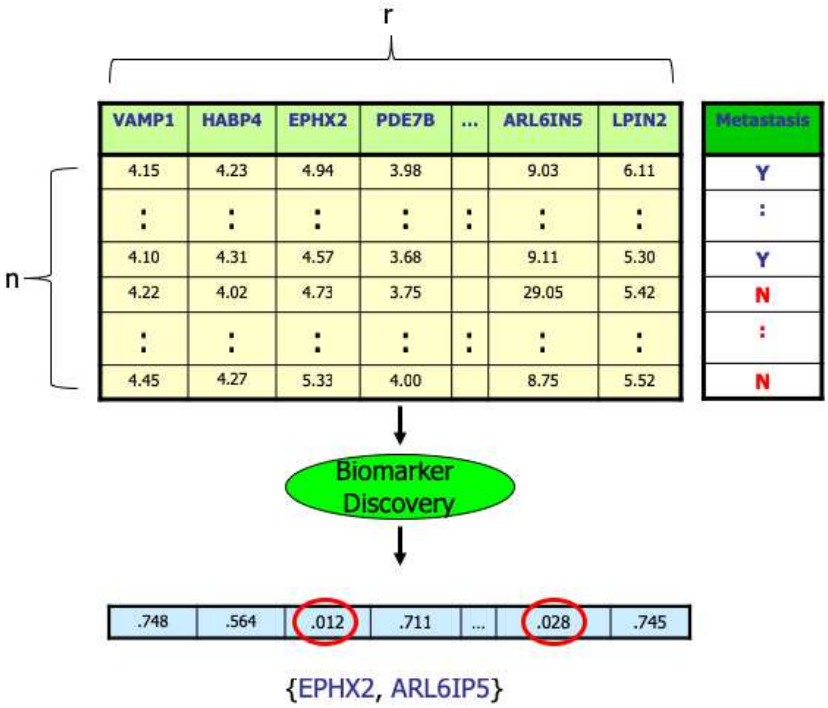

Figure 1.1: Data matrix, showing t-test p-values for each (shown) feature for the GSE 7390 dataset [14], wrt the group label (here "Metastasis" for breast cancer); the circled features, with $p < 0.05$, are (purported) biomarkers.

There are many possible reasons for this. (1) Each study should consider the *same well-defined "distribution" over instances* – *e.g.*, over the same distribution of ages and genders, etc. If the study is distinguishing case from control, then the two sub-populations should differ in only this single characteristic, but should otherwise be the same – perhaps pre-treatment women over sixty years old – and well defined label: whether each of these women developed breast cancer within 5 years. Unfortunately, matching cases and controls over all possible covariates is often not achievable. (2) A second issue is defining exactly what "*reproducible*" means – *e.g.*, is it a property of a *specific biomarker*, or of *a set of biomarkers*? In either case, what is the best objective measure to use? This is especially problematic when dealing with multifactorial diseases, where the label corresponds to a disjunction over many sub-diseases. (3) A final important issue is the *sample size*: many studies have relatively few instances, which increase the chance of finding both false negatives and false positives.

Our analysis assumes the researchers have addressed the first "sampling selection" issue (1), by running carefully designed, well-specified studies. Further, we also assume that there is no uncertainty in the labels with respect to

its clinical or biological definition. We will provide a precise measure of reproducibility (2), as well as some specific implementations, and show empirically how this varies with sample size (3).

This dissertation will focus on the simplest type of biomarkers: single stand-alone features. Note each feature could be a pre-defined combination of single features (*e.g.*, the average expression values of the genes associated with a pre-defined signalling pathway – see gene enrichment [53][5]), but we are not considering *learning* combinations. We will assume there is a *Biomarker Discovery* process, $\text{BD}(\cdot)$, that, given a labeled data matrix of n instances over a set of r features, identifies a subset of those features, which it returns as a set of (proposed) biomarkers; see Figure 1.1.[6]

As noted above, we are not considering approaches that validate proposed biomarkers based on further (or prior) biological studies, nor on downstream learned predictors. Instead, we consider computational (not biological) ways to validate such feature sets – following the intuition that biomarkers should (at least) be consistent.

In particular we define the *Reproducibility Score* $\text{RS}(D, \text{BD})$ that quantifies the "reproducibility" of the set of proposed biomarkers $\text{BD}(D)$, produced by running BD over the size-n labeled dataset D: *viz.*,

> the average Jaccard score between these proposed biomarkers $\text{BD}(D)$,
> and those produced by running the same BD process over another size-n dataset drawn from the same distribution. (1.1)

(We give a formal definition in Section 2.1.)

This dissertation presents a framework that describes the Reproducibility Score and the challenges of estimating this measure and defines three approximations for RS: an overbound, an underbound and a middle-value form, then provide empirical tests over many datasets – microarray and RNAseq data (with real values) and SNP data (discrete values) – and focusing only on t-test as the main Biomarker Discovery process $\text{BD}(\cdot)$, to confirm the effectiveness of these approximations. Researchers can use this framework, as a first step, to estimate the reproducibility of the potential results of their biomarker discovery study. A low RS suggests that these biomarkers might not be accurate which may mean that the size of the dataset used is too small or the dataset is

[5] As a subtle point: these features could be based on information *previously learned* from another dataset, or could be re-encodings of the current data, perhaps based on Principle Component Analysis [27].

[6] Note this is a single step. Some more modern GWAStudies involve many phases – typically using one phase to reduce $\approx 10^6$ features to a few thousand based on one dataset, and then using a second dataset to reduce those features to a sub-subset, etc [46]. Here, our analysis is relevant to any one of these phases; see Figure 1.1. Many studies regressed out covariates before finding biomarkers, we assume this has happened and our analysis takes those regressed out values.

too heterogeneous or perhaps the BD(·) algorithm used is not suited for this dataset.

We first close this section by motivating the need for an objective measure for evaluating the quality of a set of biomarkers (Section 1.1), then providing a short overview of why it can be difficult to find biomarkers, in general (Section 1.2) and finally overviewing some earlier studies that discuss the issue of reproducibility in biomarker discovery and/or provide approaches that could be beneficial when dealing with such problems (Section 1.3). After this, Chapter 2 first provides a formal description of the problem then explains the three approximations, and associated algorithms, that we suggest for the *Reproducibility Score*, and then describes some of the standard BD(·) algorithms. Chapter 3 summarizes our empirical experiments and the datasets used, and reports the results of the empirical study over many datasets, to show that our system works effectively – *i.e.*, that our assumptions hold true.

Chapter 4 summarizes some future work and the contributions of this dissertation. The appendices provide auxiliary information: discussion of how Biomarker Discovery differs from standard (supervised) Machine Learning, and results from other empirical studies, which explore how the RS varies with the type of FDR correction used (including "none"), the p-value threshold and the number of iterations of the approximation algorithms.

1.1 Motivation for Evaluating Biomarker Sets

To motivate the need for evaluation for association studies, consider first *predictive studies*, which use a labeled dataset, like the one shown at the top of Figure 1.1, to produce a predictive model (perhaps a decision tree, or a linear classifier) that can be used to classify future instances – here into two classes (there Y vs N). Of course, in addition to the learned classifier, the researchers will also compute *a meaningful estimate of its quality – i.e.*, of the accuracy (or AUROC, or Kappa Score, or . . .) of this classifier on an independent held-out set [57], or the k-fold cross-validation results over the training sample – perhaps "78 ± 2%" accuracy. There are (at least) three obvious things to do with this evaluation score: (1) Researchers can use this score when comparing different learning algorithms, seeking the learned classifier that gives the best average accuracy.[7] (2) If the score of this best algorithm is low (say only 51% accuracy on a balanced, binary dataset), the researchers will probably decide there was not sufficient signal in the data, or the dataset was too small to reveal it. (3) Finally, if the researchers decide to disseminate that learned classifier (*e.g.*, in a publication), they will of course announce that estimated score, along with the learned classifier. Note this can serve as a falsifiability claim: if future users run that learned model on a dataset, from the same patient distribution, they should expect to find its accuracy is at least $78 - 1.96 \times 2\%$.

[7] Of course, they have to be careful to avoid overfitting; see Witten & Frank [57].

Observing an accuracy that is much below that score, over a dataset from this same distribution, strongly suggests that the presented model is wrong.[8]

By contrast, many association studies report only a set of purported biomarkers, but provide no (estimated) score. Given that biomarkers should be reproducible (note the number of meta-reviews that claim that a set of biomarkers is problematic if they are not reproduced in subsequent studies [18], [33], [61]), we propose evaluating a biomarker set with its reproducibility score; see Equation 1.1. An accurate estimate of this score can help in the three ways discussed above, for predictive studies:(1) Researchers can compare various different "comparable" $\text{BD}(\cdot)$ algorithms, to see which produces the biomarker set that is most reproducible. This "comparable" corresponds to standard practice, where we only consider discovery tools that impose some criterion, such as the same p-value, or only considering features that exhibit a minimum fold-change. This automatically means we would not be comparing discovery tools that use $p = 0.05$ with others that use $p = 1.0$.[9] (We will see that FDR-correction, while useful in removing false-positives, can be detrimental to the goal of producing reproducible biomarkers; similarly, there is no reason to insist on p <0.05 for the statistic test used.) This could also help answer the question of whether we should use some other criterion (such as fold-change; see Section 1.3) as well as p-value, for determining the best set of biomarkers. Indeed, this type of analysis might help debug a problematic $\text{BD}(\cdot)$ algorithm. (2) A low RS score, for the best $\text{BD}(\cdot)$ algorithm, suggests that few of these purported biomarkers will be found in another dataset, which argues these purported biomarkers might not be accurate – which might argue for not using such a small dataset, etc. (3) Finally, there are many meta-reviews that note that different studies find different sets of biomarkers and question whether the techniques used are to blame – *e.g.*, [19], [29], [36]. One way to address this concern is to require that each published paper include both the purported set of biomarkers, and also an estimate of its reproducibility score, RS. The same way a prediction study's "5 fold cross validation" accuracy tells the reader how accurate the classification model should be on new data, this reproducibility score would similarly tell the reader whether to expect another study, on a similar dataset, will find many of the same biomarkers. A low RS score suggests that few of these purported biomarkers will appear in another dataset, which argues they might not be strongly associated with the label (think "disease"). Reviewers, and other critics, might then argue that these results might not be meaningful – either because the dataset is too small or too heterogeneous or the $\text{BD}(\cdot)$ technique used is problematic. Note that we should view this RS test as *necessary* for considering a proposed model, but not sufficient – *i.e.*, it might *rule-out* a proposed discovery model, but should

[8] Of course, this assumes that the evaluation of the model's performance is done correctly; and even then, this claim is only with 95% confidence. It is still suggestive of a problem.

[9] A tool that uses $p = 1.0$ – such as $\text{BD}_{t,1.0,BH}$, defined below – would return all features for any dataset, and so necessarily have a Jaccard of 1.0.

not be enough to *rule-in* a model.

Despite the tight coupling showing that evaluation goals from Supervised Machine Learning can apply to our Biomarker Discovery task, there are several significant differences between these two tasks. In general, a predictive model provides some information about an individual – eg, whether she has some disease. By contrast, an association study identifies features, with the prediction that they will each exhibit some population difference wrt a dataset of many individuals. Also, it is relatively easy to evaluate the quality of a learned predictive model, by running that predictor on a held-out set of instances. By contrast, there is no direct way to determine if a purported biomarker is correct. This is why we, instead, look for "consistency" of a set of biomarker discovery tools. That is, we hope that these discovered feature sets have low variance. (Note that they can have high-bias – eg, if they all set $p = 1$, then each discoverer will return all features; this will have low variance, but presumably high bias.)

1.2 Why is Biomarker Discovery so difficult?

Above we observed that a set of biomarkers might not be reproducible across different studies. This might be due to the difficulty of identifying biomarkers from a small sample. Another challenge is based on the nature of univariate biomarkers in general: When the target is a complex disease or condition, a single feature is not sufficient to accurately explain the outcome; indeed there are situations where a *combination* of features are important, but each of the component features, by itself, is completely irrelevant, and so would not qualify as a biomarker.

As an example, consider a baby *in utero*, and note that its health may be completely uncorrelated with the Rh blood type of its mother, MRh $\in \{+, -\}$, and is also completely uncorrelated with the father's Rh factor, FRh – which means neither MRh nor FRh could be a biomarker. However, suppose finding these blood factors are different MHr$\neq$FHr, increases the baby's risk. Assuming balanced sampling (with an equal number of MRh=+ and MRh=−, and similarly for FRh), this means an effective predictive model would need to include both features, even though neither is a biomarker. We typically assume that a feature either increases the risk of a disease in all situations, or always decreases that risk. This in-utero-baby example shows this is not always the case: We see that MRh=+ can sometimes increase the risk (when FRh=−), and other times, decrease the risk (when FRh=+). Hence, a simple linear combination of feature values might not always be appropriate.

While this is an extreme situation – where each feature is completely irrelevant by itself – it is relatively common for a disease to be associated with many minor features; here again, it is possible that none of the features, by itself, shows sufficient class distinction to qualify as a biomarker. This also happens when the class is inherently heterogeneous – *e.g.*, "headache" can be

based on various phenomena, including ischemic stroke, dehydration, migraine, etc., each with various different factors. This is believed to happen with essentially all complex genetic disorders, especially when underlying pathologies are not known.

These situations argue that a *panel* of features can sometimes be more appropriate than individual features. If the model starts with a pre-defined combination of a set of features – *e.g.*, a simple average of a specific set of gene expression values, or the number of heterozygous settings in a specific set of SNPs – then we can view that combination as a (super-)feature, and let it be a column in the matrix of Figure 1.1; the analysis described in the dissertation still apply. Note, however, that here we assume this super-feature construction is known initially, and in particular, this dissertation is *not* exploring ways to find these features – *i.e.*, it is not describing machine learning tools for producing new super-features. We are also not considering multivariate approaches, where one feature can implicitly condition on other features simultaneously – *e.g.*, multiple regression models.

1.3 Related Work

There have been many pairs of studies that have each produced biomarkers for the same disease or condition, but found little or no overlap between the two lists of purported biomarkers. Many papers have discussed this issue – some describing this problem in general [18], [19], [61], and others exploring specific examples[33], [62]. These papers suggest different causes for the problem, such as the heterogeneous biological variations in some datasets [18], [61] or problems in the methods used that may lead to non-reproducible results [22], [48].

In particular, Zhang *et al.* [61] challenge the claim that the non-reproducibility problem in microarray studies is due to poor quality of microarray technology, by showing that inconsistencies occur even between technical replicates of the same dataset. They also show that heterogeneity in cancer pathology would further reduce reproducibility.

Ein-Dor *et al.* [18] also show the inconsistencies between the results of subsamples of a single dataset, demonstrating that the set of (gene) biomarkers discovered is not unique. They explain that there are many genes correlated with the group labels, but the empirical correlations change for different (sub)samples of instances. These two papers motivate our need for tools that can effectively bound the reproducibility – such as the ones presented here.

Several projects [19], [22], [48] have attempted to formally analyse this problem. Ein-Dor *et al.* [19] describe a method, probably approximately correct (PAC) sorting, that estimates the minimum number of instances needed for a desired level of reproducibility. As an example, this worst-case analysis proves that, to guarantee a 50% overlap between different gene lists for breast cancer, each dataset needs to include at least several thousand patients. This

suggests poor repeatability results when using small sample sizes, which is consistent with our results for datasets with smaller sample sizes; see Chapter 3, especially Figure 2.3.

The goal of the MicroArray Quality Control (MAQC) project [48] was to address the problems and uncertainties about the microarray technology that were caused by the observation that different studies (of the same phenotype) often found very different biomarkers. They suggest that the common approach of using just t-test p-values (specifically stringent p-values) can lead to poor reproducibility, which motivated them to consider methods like fold-change ranking with a non-stringent p cutoff, which they demonstrate leads to more reproducible gene sets. In a follow-up, Guo *et al.* [22] found similar results by using the same procedures for another dataset.

However, Klebanov *et al.* [30] later show that these MAQC project results do not prove that using t-tests is necessarily unsuitable – *i.e.*, just because another method (here fold-change) can generate more reproducible results, does not mean that it is performing better; as an extreme, the algorithm that declares every gene is a biomarker, is completely reproducible; see Footnote 9. They demonstrate these points by using a set of simulation studies (where they know the "true biomarkers"), and use either t-test or fold-change to propose potential biomarkers. These studies found that the t-test approach performed much better than the fold-change, in terms of recall. These results motivated us to use the t-test approach (rather than fold-change) as our main BD algorithm – which we use for all of our empirical experiments.

Chapter 2

Materials and Methods

2.1 Formal Description

As suggested by Figure 1.1, a "Biomarker Discovery" algorithm, $\text{BD}(\cdot)$, takes as input a dataset D of n instances, each described by r features $F = \{f_1, \ldots, f_r\}$ and labeled with a binary class, and returns a subset of features $F' \subset F$ of purported biomarkers, where each $f \in F'$ shows a class difference. That is, letting x_i^j be the value of the i^{th} feature of the j^{th} instance, and $\ell^{(j)}$ be the label of the j^{th} instance (which is either Y or N), the set $\{x_i^j \,|\, \ell^{(j)} = Y\}$ of values of the i^{th} feature of the diseased individuals, is significantly different from the values of that feature over the healthy individuals $\{x_i^j \,|\, \ell^{(j)} = N\}$. We will assume that these $\{\, x_i^j \,\}_{i,j}$ values are either all continuous (such as height, or the expression value of a gene), or all discrete (think gender, or the genotype of a SNP). (Section 2.2 below will describe several such $\text{BD}(\cdot)$s.)

As noted in Equation 1.1, the *Reproducibility Score* $\text{RS}(\ D, \text{BD}\ )$ quantifies the "reproducibility" of the set of proposed biomarkers $\text{BD}(\ D\ )$, produced by running BD over the size-n labeled dataset D. Here, we assume that the values of each feature $x_\cdot^j$, for each label c, are generated independently from a fixed distribution (*i.e.*, "iid") $p_{j,c}(\cdot) = p(\,\bar{x}^{(j)} \,|\, \ell^{(j)} = c\,)$. Note these are just the marginals; we do not assume that the various features are independent from one another, – *i.e.*, this does not necessarily correspond to Naive Bayes [57]. We will view $p(\cdot) = \{\, p_{j,c}(\cdot) \,\}_{j,c}$ as the set of these $2 \times r$ different distributions, and let $p^n(\cdot)$ be the distribution for drawing n instances, iid, from this set of distributions. Then

$$RS^*(\, p(\,\cdot\,),\, n,\ \text{BD}(\,\cdot\,)\,) \quad = \quad E_{D',D" \sim p^n(\cdot)}[\ J(\,\text{BD}(\text{D}'),\, \text{BD}(\text{D}")\,)] \tag{2.1}$$

where the Jaccard score of two sets

$$J(\,A,\,B\,) \quad = \quad \frac{|A \cap B|}{|A \cup B|} \tag{2.2}$$

is the ratio of the intersection to the union of these sets – hence $J(\,A,\,B\,)$ ranges from 0 to 1, and is 1 iff $A = B$, and is 0 iff these sets are disjoint.

(Appendix A.3 discusses another measure that is sometimes used to measure the reproducibility of a set of biomarkers.)

Of course, we do not know $p(\cdot)$, and so we use the empirical distribution, based on the dataset D – call it $\widehat{p_D}(\cdot)$ – to produce:

$$\text{RS}(\, D,\, \text{BD}(\cdot)\,) \quad = \quad RS^*(\,\widehat{p_D}(\cdot),\, |D|,\ \text{BD}(\cdot)\,) \tag{2.3}$$

which measure the reproducibility of the biomarker set $\text{BD}(\,\text{D}\,)$. *N.b.*, this Reproducibility Score deals with the *set* of biomarkers that is produced by the $\text{BD}(\cdot)$ function, and not any single specific biomarkers.

Of course, Equation 2.3 suggests the obvious bootstrap sampling algorithm [17]. Empirically, however, we found that it did not perform well – motivating the algorithms described in Section 2.3.

2.2 Biomarker Discovery Algorithms: $\text{BD}(\cdot)$

The previous sections discussed how to evaluate the result of applying some biomarker discovery algorithm $\text{BD}(\cdot)$ on a labeled dataset, and provided some approximations here. This section describes some of the standard biomarker discovery algorithms.

Initially there are two types of datasets, depending on whether its feature values (the x_i^j mentioned above) are continuous or discrete. However, for datasets with categorical values – SNPs in our analysis – we use a simple preprocessing step, which precedes all the $\text{BD}(\cdot)$ algorithms described here, to convert each categorical value to a real number, allowing us to view each such dataset as one with continuous values. In particular, we convert each SNP feature, which ranges over the values { AA, Ab, bb }, to the real-values { 0, 1, 2 }, corresponding to the number of minor alleles ("b") in the genotype.

Here we assume that the real values of each feature (column) follow a normal distribution, which might be different for the different classes, and so we use a t-test (independent two-sample t-test) for all of our empirical experiments:

$$\begin{aligned} \text{t} &= \frac{\bar{X_Y} - \bar{X_N}}{s_p \cdot \sqrt{\frac{1}{n_Y} + \frac{1}{n_N}}} \\ s_p &= \sqrt{\frac{(n_Y - 1)\,\bar{s}_Y^2 + (n_N - 1)\,\bar{s}_N^2}{n_Y + n_N - 2}}. \end{aligned} \tag{2.4}$$

where n_Y and n_N are the sample sizes of instances with label Y and label N, respectively, with empirical means $\bar{X_N}$ and $\bar{X_Y}$ and empirical variances $\bar{s}_Y^2$ and $\bar{s}_N^2$.

Notice the biomarker discovery process is basically performing one statistical test for each of a large number of features – often tens-of-thousands,

or more! This has motivated many researchers to seek ways to reduce the chance of false discoveries – often by using some FDR (False Discovery Rate) correction. Our studies focus on the Benjamini/Hochberg approach [5].

We refer to the resulting tool as $\text{BD}_{t,0.05,BH}(\cdot)$, where the t in the subscript refers to the 2-sided t-test, the 0.05 for the p-value used, and BH to the Benjamini/Hochberg correction. This notation makes it easy to consider many variants – *e.g.*, adjusting the p-value used for the statistical test, whether it is applying another multiple testing correction, or none, etc.

2.3 Algorithms that Approximate the Reproducibility Score

As we have the dataset D with n labeled instances, we can directly compute $\text{BD}(\ D\ )$. To compute $\text{RS}(\ D,\ \text{BD}(\cdot)\ )$, we need to produce one (or more) similar datasets D', each with n instances drawn from the same (implicit) distribution $p_D(\cdot)$ that generated D, but which is presumably disjoint from D. While we do not have such D''s, and so cannot directly compute the Reproducibility Score, we can compute an overbound, an underbound and a middle-value estimate of $\text{RS}(\ D,\ \text{BD}(\cdot)\ )$.

2.3.1 Overbound

The $\text{oRS}(\ D, \text{BD}(\cdot), k)$ procedure produces (an estimate of) an overbound of $\text{RS}(\ D,\ \text{BD}(\cdot)\ )$, by making it *easier* for a feature to be selected to be in both purported biomarker sets. oRS first defines a size-$2n$ dataset DD that contains two copies of each instance in D, of course with the same label both times. It then randomly "partitions" DD into two size-n datasets $D1$ and $D2$, balanced by label.[1] It then runs $\text{BD}(\cdot)$ on each, to produce two biomarker sets, then computes the Jaccard score for this pair of biomarker sets: $\text{J}(\ \text{BD}(\text{D1}),\ \text{BD}(\text{D2})\ )$. As we expect $D1$ to overlap with $D2$, it is relatively likely that any $D1$-biomarker will also be a $D2$-biomarker (more likely than if $D1$ was disjoint from $D2$), which means we expect the associated Jaccard score to be higher. This follows from the observation that, as two datasets have more common elements, we expect the number of biomarkers common to two datasets, to increase – *i.e.*,

$$\begin{array}{ccccc} \text{if} & A_1, A_2,\ B_1, B_2 \sim p^s(\cdot) \text{ and} & & & \\ & |A1 \cap A2| & \text{is larger than} & & |B1 \cap B2|, \\ \text{then we expect} & |\text{BD}(\text{A1}) \ \cap\ \text{BD}(\text{A2})| & \text{will be larger than} & & |\text{BD}(\text{B1}) \ \cap\ \text{BD}(\text{B2})| \end{array} \tag{2.5}$$

[1] This "partitioning" is wrt the *list* of elements, which can include duplicates. Also, this process actually keeps the datasets "balanced", in terms of labels – this requires partitioning DD into DD^+ and DD^-, where DD^+ are the cases and DD^- the controls. We then form $D1^+$ by randomly drawing 1/2 of DD^+, and $D1^-$ by randomly drawing 1/2 of DD^-, then forming $D1 = D1^+ \cup D1^-$. See Figure 2.1.

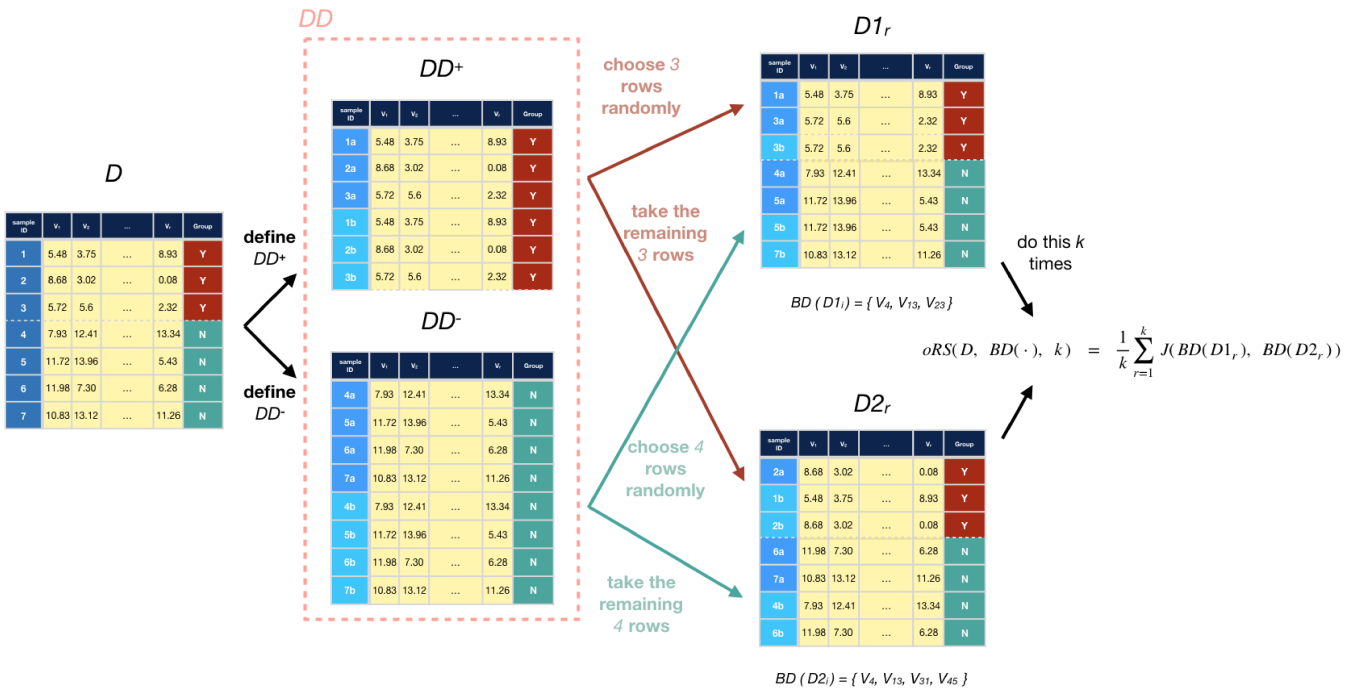

Figure 2.1: Diagram showing the process of generating pairs of subsets for a dataset D and then computing the $\text{oRS}(\, D, \text{BD}(\,\cdot\,), k)$.

ceteris paribus. Our $\text{oRS}(\, D, \text{BD}(\,\cdot\,), k)$ algorithm actually computes k dataset-pairs $\{[D1_r,\ D2_r]\}_{r=1..k}$, whose list-union is DD (*i.e.*, $D1_r + D2_r = DD$ for each r), and returns the average

$$\text{oRS}(\, D, \text{BD}(\,\cdot\,), k) \quad = \quad \frac{1}{k} \sum_{r=1}^{k} J(\, \text{BD}(D1_r),\ \text{BD}(D2_r)\,) \ . \tag{2.6}$$

It is easy to relate this approach to RS^* (Equation 2.1), as each $D1_r$ and $D2_r$ are drawn from $p_D(\,\cdot\,)$. They are not quite bootstrap samples as this approach means each instance will occur exactly twice in the (list)union of $D1_r + D2_r$, while in a standard bootstrap sample, an instance might occur many times in each individual drawn sample.[2]

[2]Some quick observations: (1) We expect 25% of the instances to be duplicated in any given dataset. This is the minimum amount of resampling required to produce size-$2n$ dataset DD. (2) In general, we expect the number of biomarkers common to two datasets, to increase as this pair of datasets includes the same instances more number of times – which would happen in bootstrap sampling. (Note that we also tried that approach, but found its associated scores to be too generous – they were higher than our current overbound scores, oRS, for all datasets.) This is why we are instead using our "doubling approach", as it produces values that are smaller, but still remains an overbound, as desired.

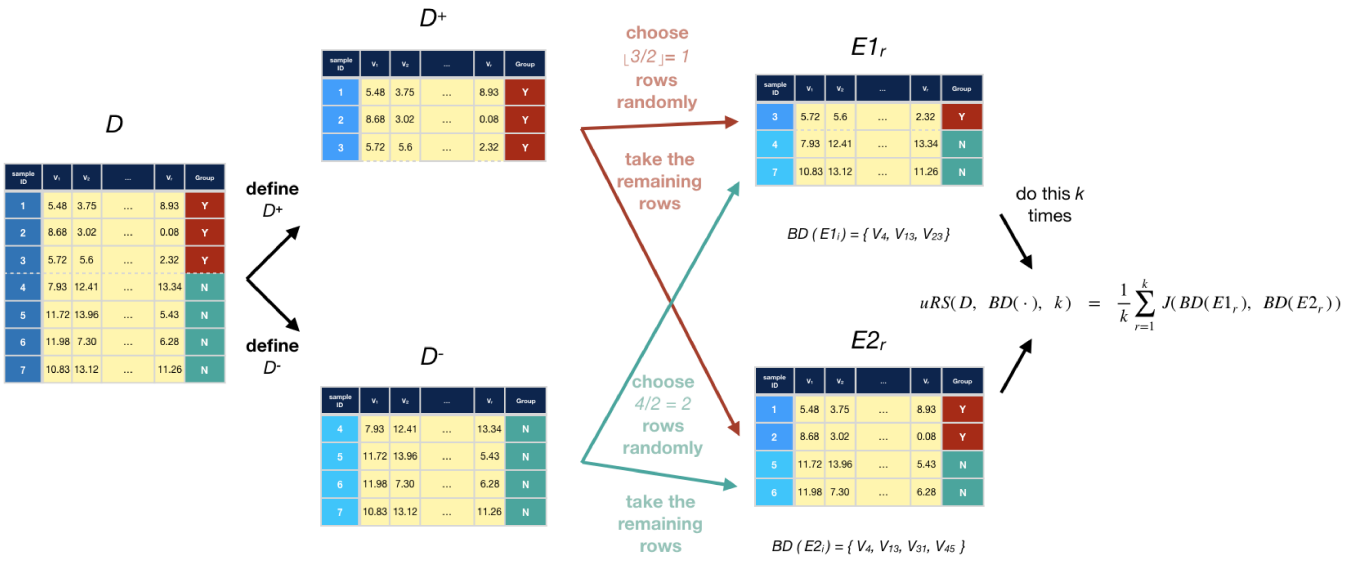

Figure 2.2: Diagram showing the process of generating pairs of subsets for a dataset D and then computing the uRS$(D, \text{BD}(\cdot), k)$.

2.3.2 Underbound

The uRS$(D, \text{BD}(\cdot), k)$ procedure produces (an estimate of) an underbound of RS$(D, \text{BD}(\cdot))$, by making it *harder* for a feature to be selected to be in both purported biomarker sets. First, observe that as n increases, we expect the statistical estimates to be more accurate, and in particular, statistical tests for differences between the two classes will be correct more often. Hence, a statistical test will better identify the "true" biomarkers F^* from an n-element dataset $D^{(n)}$, versus from an $n/2$-element dataset $D^{(n/2)}$. Now consider two n-element datasets $D1^{(n)}$ and $D2^{(n)}$, and also two $n/2$-element datasets $E1^{(n/2)}$ and $E2^{(n/2)}$. As $\text{BD}(D1^{(n)})$ and $\text{BD}(D2^{(n)})$ are each closer to F^* than $\text{BD}(E1^{(n/2)})$ and $\text{BD}(E2^{(n/2)})$, we expect $\text{BD}(D1^{(n)})$ and $\text{BD}(D2^{(n)})$ to be closer to each other, than $\text{BD}(E1^{(n/2)})$ and $\text{BD}(E2^{(n/2)})$, which means we expect J($\text{BD}(D1^{(n)})$, $\text{BD}(D2^{(n)})$) to be larger than J($\text{BD}(E1^{(n/2)})$, $\text{BD}(E2^{(n/2)})$) – *i.e.*, we expect $RS^*(p(\cdot), n, \text{BD}(\cdot))$ to be larger than $RS^*(p(\cdot), n/2, \text{BD}(\cdot))$. (Figure 2.3 shows this idea in general: given that

$$RS^*(p(\cdot), s, \text{BD}(\cdot)) \approx \hat{E}^{(k)}[J(\text{BD}(\text{D1}^{(s)}), \text{BD}(\text{D2}^{(s)}))] \tag{2.7}$$

we see that the RS^* score increases with the size s of the dataset – *i.e.*,

$$\begin{array}{lcccc} \text{if} & A_1, A_2 \sim p^{sA}(\cdot),\ B_1, B_2 \sim p^{sB}(\cdot) \text{ and} & & & \\ & sA = |A1| = |A2| & \text{is larger than} & sB = |B1| = |B2|, & \\ \text{then we expect} & |\text{BD}(\text{A1}) \cap \text{BD}(\text{A2})| & \text{will be larger than} & |\text{BD}(\text{B1}) \cap \text{BD}(\text{B2})| & \end{array} \tag{2.8}$$

Our uRS$(D, \text{BD}(\cdot), k)$ algorithm first partitions D into two disjoint $n/2$-instance subsets, $E1$ and $E2$, with balanced labels. It then computes J($\text{BD}(E1)$, $\text{BD}(E2)$) which, by the argument above, should be an underbound

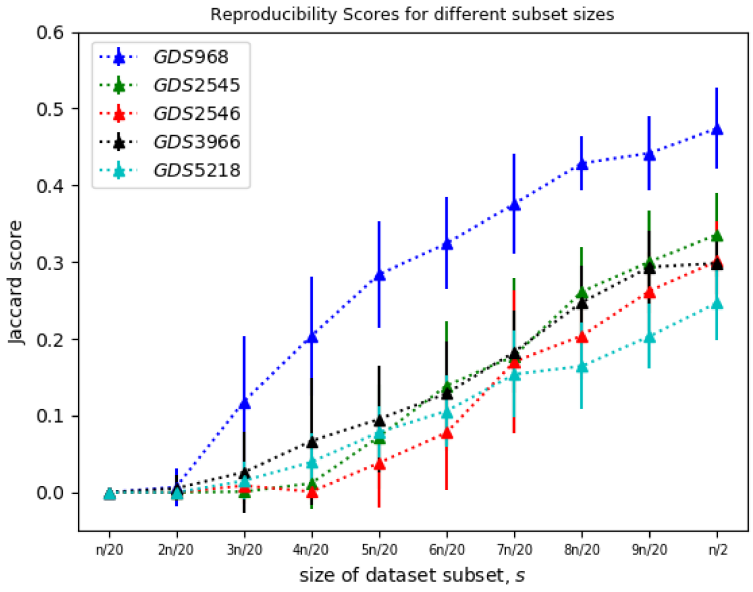

Figure 2.3: For each of the 5 datasets D , each point shows the average ($\pm$ sd) Jaccard $\hat{E}^{(k)}[\, J(\,\text{BD}(\text{D1}^{(s)}),\, \text{BD}(\text{D2}^{(s)})\,)\,]$ over $k = 20$ pairs $[D1^{(s)},\ D2^{(s)}]$ of disjoint size-s subsets of D, shown as a fraction of n, which is the size of the original dataset. Note the x-axis can only go to $n/2$, and we are using the standard $\text{BD}_{t,0.05,BH}$.

on RS(D, BD($\cdot$)). It actually does this partitioning k times, producing k different $\{\,[E1_r,\ E2_r]\,\}_{r=1..k}$ dataset pairs (each pair being disjoint, and each dataset of size $n/2$) and returns the average; see Figure 2.2.

$$\text{uRS}(\,D, \text{BD}(\,\cdot\,), k) \quad = \quad \frac{1}{k}\sum_{r=1}^{k} J(\,\text{BD}(\text{E1}_r),\, \text{BD}(\text{E2}_r)\,)\ . \tag{2.9}$$

2.3.3 Middle-Value Approximation

Recall our goal is estimating the overlap between the biomarkers of two size-n datasets; one reason why uRS is an underbound is that it uses only size-$n/2$ datasets. This motivates us to extend uRS to deal with size-n datasets. This leads us to the middle-value which is very similar to the underbound uRS, as it also first partitions D into two disjoint, balanced subsets, $E1$ and $E2$. Here, however, it "extends" each subset: That is, if D contained 100 instances, then $E1$ and $E2$ would each have 50 instances. The mRS routine, however, essentially uses the 100-instance $ED1$, which has the same empirical variance and empirical mean as $E1$ but double the sample-size, and $ED2$ which has

the same empirical variance and mean as $E2$ and double the sample-size.[3] As with uRS, we consider k different partitions, etc. Hence,

$$\text{mRS}(\, D, \text{BD}(\,\cdot\,), k) \quad = \quad \frac{1}{k} \sum_{r=1}^{k} J(\, \text{BD}(\, ED1_r\,),\, \text{BD}(\, ED2_r\,)\,)\ . \qquad (2.10)$$

As mRS uses datasets that are twice as large as the ones used by uRS, we anticipate $\text{mRS}(\, D, \text{BD}(\,\cdot\,), k) \geq \text{uRS}(\, D, \text{BD}(\,\cdot\,), k)$ (see Equation 2.8), and as mRS's pair of datasets are disjoint, while oRS's pair (typically) are not, we anticipate that $\text{mRS}(\, D, \text{BD}(\,\cdot\,), k) \leq \text{oRS}(\, D, \text{BD}(\,\cdot\,), k)$ (see Equation 2.5).

[3] If BD is using a t-test (Equation 2.4) – such as $\text{BD}_{t,0.05,BH}$ – then that test would basically involve simply doubling the values of n_N and n_Y, so $n_N + n_Y = 100$ rather than 50.

Chapter 3

Empirical Study over Various Datasets and Results

3.1 Empirical Study

There are now many publicly-available datasets that have been used in association studies. Here, we use them to (1) Better understand what Jaccard scores are typical, for a range of standard $\text{BD}(\cdot)$ algorithms; (2) Determine whether our predictions match the results of earlier meta-analyses; and (3) Determine if our approximations are meaningful – *i.e.*, if (for large values of k):

$$\text{uRS}(D, \text{BD}(\cdot), k) \leq \text{RS}(D, \text{BD}(\cdot)) \tag{3.1}$$

$$\text{oRS}(D, \text{BD}(\cdot), k) \geq \text{RS}(D, \text{BD}(\cdot)) \tag{3.2}$$

$$\text{mRS}(D, \text{BD}(\cdot), k) \approx \text{RS}(D, \text{BD}(\cdot)) \tag{3.3}$$

The next section will explicitly discuss (1) and (2). It is trickier to deal with (3): Given only a single dataset D of size-n, we cannot compute, nor even estimate, the true value of $\text{RS}(D, \text{BD}(\cdot))$. However, we can estimate $\text{RS}(D^{(n/2)}, \text{BD}(\cdot))$, where $D^{(n/2)}$ is a size-$n/2$ subset of D. In fact, $\text{uRS}(D, \text{BD}(\cdot), \text{k})$ is a meaningful estimate of $\text{RS}(D^{(n/2)}, \text{BD}(\cdot))$; below we will use

$$\widehat{RS}(D^{(n/2)}, \text{BD}(\cdot), k) = \text{uRS}(D, \text{BD}(\cdot), k) \tag{3.4}$$

We will then compare this $\widehat{RS}(D^{(n/2)}, \text{BD}(\cdot), k)$ against $\text{uRS}(D^{(n/2)}, \text{BD}(\cdot), k)$, $\text{oRS}(D^{(n/2)}, \text{BD}(\cdot), k)$ and $\text{mRS}(D^{(n/2)}, \text{BD}(\cdot), k)$, to see whether the relations of Equation 3.1, 3.2 and 3.3 all hold, wrt various size-$n/2$ subsets, $D^{(n/2)}$.

We can in fact do this for any size-s subset $D^{(s)}$ of D where $s \leq n/2$. Here, we need a set of pairs of disjoint label-balanced subsets D', $D'' \subset D$ where $|D'| = |D''| = s$ and $D' \cap D'' = \{\}$. For a fixed dataset D, and specified number $k \in \mathcal{Z}$, we can then plot these $\widehat{RS}(D^{(s)}, \text{BD}(\cdot), k)$ values along with $\text{oRS}(D^{(s)}, \text{BD}(\cdot), k)$, $\text{uRS}(D^{(s)}, \text{BD}(\cdot), k)$, and $\text{mRS}(D^{(s)}, \text{BD}(\cdot), k)$ as a function of s, to see their behaviour; see Figure 3.1.

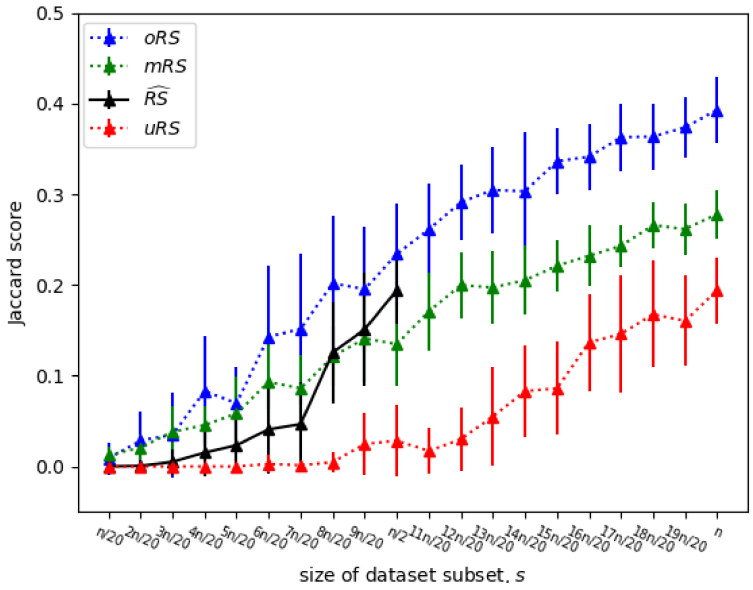

Figure 3.1: Showing how the approximations relate to one another, and scale with the size s of the dataset. Here we are using subsets of the Metabric dataset, with n=1654. We observed the same behavior for all datasets, *i.e.*, oRS $\geq$ $\widehat{RS}$ $\approx$mRS $\geq$ uRS, as the subset size s increases.

We explored our approximations over 25 different datasets, consisting of 16 microarray datasets and 2 RNAseq datasets with continuous data (see Table 3.1). This first set includes 4 of the gene expression datasets discussed in the Zou *et al.* [62] meta-analysis – each describing metastatic versus non-metastatic breast primary cancer subjects[1] – to see if our method is consistent with their empirical results. We also included 11 other relatively-small gene expression datasets (from 19 to 187 instances), focusing on human studies that had a binary class label from the GEO repository. To explore how our tools scale with size, we also included 3 other relatively large datasets, with 532 to 1654 instances. As these were survival datasets, we set the binary label based on the median survival time (removing any instance who was censored before that median time). In addition to these 4+11+3 = 18 gene expression datasets (with real-valued entries), we also include 7 SNP datasets (from 39 to 164 instances), with discrete values, also selected from human studies with binary class labels; see Table 3.3.

[1] We were not able to replicate the results reported for the 5^{th} dataset using our BD algorithm, and so had to exclude that data.

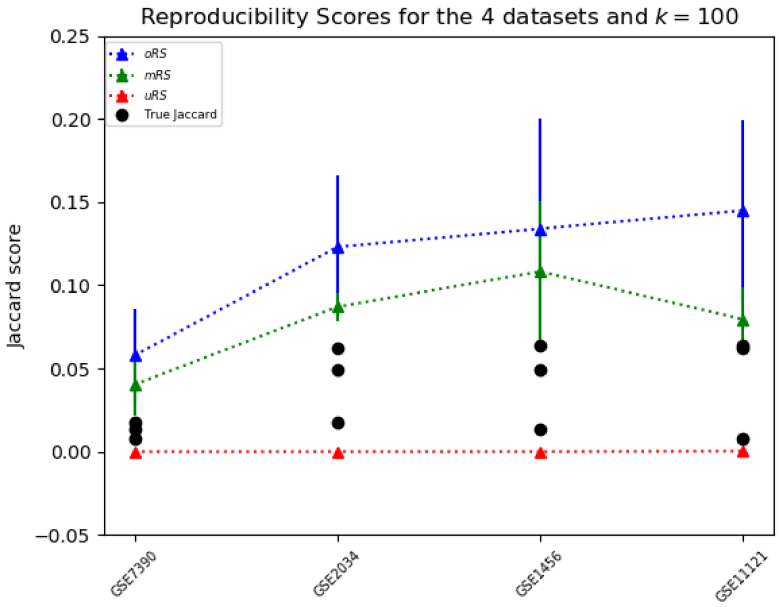

Figure 3.2: Under-bound and over-bound for the 4 datasets and also the true Jaccard score for each pair – 3 numbers for each dataset, shown by black circles.

3.2 Results

We run our suite of methods over 25 different datasets, including 16 microarray datasets and 2 RNAseq datasets, whose feature-values $\{x_i^j\}$ were real numbers (recall each x_i^j is the expression value of the i-th gene for the j-th subject; we log_2-transformed the values from the RNAseq datasets), and 7 were SNP datasets, with categorical entries – *i.e.*, each $x_i^j \in \{0,\ 1,\ 2\}$ is the number of minor alleles in the genotype for the i-th SNP for the j-th subject; see Tables 3.1 and 3.3. Here, we use the the standard $\text{BD}_{t,0.05,BH}(\cdot)$ biomarker discovery algorithm.

First, we analyzed the 4 datasets mentioned in the Zou *et al.* [62] meta-analysis (see the first 4 rows of Table 3.1) and computed the $\{\ \text{uRS}(\ D, \text{BD}_{t,0.05,BH}, 50)$, $\text{mRS}(\ D, \text{BD}_{t,0.05,BH}, 50)$, $\text{oRS}(\ D, \text{BD}_{t,0.05,BH}, 50)\ \}$ values for each dataset D, as well as the actual Jaccard score for biomarkers for each pair of datasets. The results, in Figure 3.2, show that the Jaccard score for each pair is well within the bounds computes by our approximations, for each of the datasets in that pair – that is, the results for $4 \times 3 = 12$ ordered-pairs of datasets are consistent with our predictions.[2]

We also analyzed the other 14 continuous datasets D, and computed the

[2] We first verified that our $\text{BD}_{t,0.05,BH}$ algorithm found the same biomarkers, as our PO scores (Equation A.1 in Appendix A.3) matched the ones they published. We can easily compute the PO scores from the Jaccard scores and numbers of biomarkers.

Table 3.1: Results for the microarray datasets when using all the instances. The first 4 entries are from the Zou *et al.* [62] meta-study. Reproducibility Scores are shown in the form of mean ± standard deviation.

Name	#instances (Majority %)	#features	#biomarkers	uRS %	mRS %	oRS %
GSE2034* [56]	286 (67%)	13245	277	0 ± 0	8.69 ± 2.44	12.6 ± 4.19
GSE11121* [47]	200 (86%)	13245	492	0.09 ± 0.2	10 ± 2.52	13.4 ± 5.89
GSE7390* [14], [38]	198 (82%)	13245	18	0 ± 0	4.8 ± 0.9	5.15 ± 2.81
GSE1456* [39]	159 (78%)	13245	443	0 ± 0	10.09 ± 2.32	13.6 ± 6.4
Metabric [12]	1654 (57%)	24368	3675	18.5 ± 3.86	26.64 ± 2.32	39.8 ± 3.56
BRCA [13]	552 (95%)	18320	2	0 ± 0	2.21 ± 1.27	2.5 ± 2.1
KIPAN [13]	532 (81%)	18271	2782	12.3 ± 4.91	24.33 ± 4.08	34.2 ± 4.74
GDS2771 [23], [51]	187 (52%)	22215	1807	0.32 ± 0.64	22.5 ± 7.75	31.68 ± 0.47
GDS2545 [11], [59]	171 (53%)	12558	4291	34.0 ± 4.09	54.1 ± 3.46	54.58 ± 0.31
GDS968 [44]	171 (53%)	5748	2506	47.5 ± 3.95	62.3 ± 3.2	63.25 ± 0.76
GDS2546 [11], [59]	167 (54%)	12553	2965	30.8 ± 4.82	49.4 ± 2.48	49.0 ± 0.55
GDS2547 [11], [59]	164 (54%)	12579	1810	23.7 ± 5.86	43.4 ± 3.37	42.66 ± 0.70
GDS4431 [3]	146 (53%)	54613	140	0 ± 0	7.51 ± 3.66	18.85 ± 0.34
GDS5218 [43]	110 (56%)	54675	10700	24.0 ± 5.09	45.0 ± 3.7	46.06 ± 0.29
GDS3966 [58]	83 (63%)	22274	6554	31.7 ± 4.71	53.7 ± 3.72	53.66 ± 0.27
GDS4185 [4], [28]	67 (58%)	22283	6	0 ± 0	7.08 ± 6.43	11.29 ± 0.83
GDS2737 [9]	37 (57%)	54675	4	0 ± 0	3.27 ± 2.2	10.58 ± 0.5
GDS4719 [20]	19 (53%)	54675	1	0 ± 0	4.63 ± 14.6	7.02 ± 0.71

oRS, uRS and mRS values when using $k = 50$ repetitions; see Figure 3.3[left]. We see that the overbound oRS is consistently larger than the middle-value approximation mRS, which in turn is larger than the underbound uRS – *i.e.*, oRS $\geq$ mRS $\geq$ uRS – as desired, Equations 3.1-3.3. Figure 3.3[right] plots the corresponding values for the $D^{(n/2)}$ datasets, that use only 1/2 of the dataset, using the same $\text{BD}(\cdot)$ algorithm and $k = 50$. It also plots the $\widehat{RS}(D^{(n/2)}, \text{BD}(\cdot), k)$ values for the datasets. Here, we see oRS $\geq \widehat{RS} \approx$ mRS $\geq$ uRS, as desired.

Finally, similar to that experiment over the 18 continuous datasets, we examined the 7 discrete datasets and produced the reproducibility scores. Figure 3.4[left] shows the scores for each of the 7 SNP datasets, demonstrating that oRS $\geq$ mRS $\geq$ uRS holds for the discrete cases as well. Figure 3.4[right] shows the scores for $D^{(n/2)}$ datasets, when using only 1/2 of the dataset and again we can see that oRS $\geq \widehat{RS} \approx$ mRS $\geq$ uRS, holds for all cases.

Appendix A provides the results of many additional empirical studies, showing how the reproduciblity scores change based on which (if any) FDR correction is used, the specific p-value used for the t-test, the number of draws k used by the various approximations, and the size of the dataset n.

Table 3.2: Results for the microarray datasets when using half of the instances. Reproducibility Scores and average number of biomarkers are shown in the form of mean ± standard deviation.

Name	Average #biomarkers	uRS %	$\widehat{RS}$ %	mRS %	oRS %
GSE2034*	41.05 ± 144.8	0 ± 0	0 ± 0	4.58 ± 4.29	6.39 ± 4.72
GSE11121*	181.8 ± 265.5	0 ± 0	0.09 ± 0.2	6.47 ± 4.74	9.48 ± 7.67
GSE7390*	1.38 ± 3.01	0 ± 0	0 ± 0	1.67 ± 1.06	1.84 ± 2.1
GSE1456*	58.38 ± 138.11	0 ± 0	0 ± 0	4.05 ± 3.94	4.5 ± 4.41
Metabric	1301.9 ± 504.80	1.97 ± 3.86	18.5 ± 3.86	13.6 ± 3.79	21.6 ± 7.3
BRCA	36.75 ± 171.15	0 ± 0	0 ± 0	1.6 ± 2.16	2.6 ± 4.2
KIPAN	694.48 ± 398.88	0.88 ± 2	12.3 ± 4.91	11.63 ± 4.03	19.3 ± 8.33
GDS2771	457.17 ± 863.64	0.09 ± 0.62	0.32 ± 0.64	19.03 ± 0.62	7.23 ± 8.48
GDS2545	2051.58 ± 494.77	6.38 ± 7.33	34.0 ± 4.09	38.51 ± 0.66	38.1 ± 5.92
GDS968	1593.67 ± 173.35	26.0 ± 5.87	47.5 ± 3.95	51.3 ± 0.79	48.6 ± 5.21
GDS2546	1266.61 ± 354.72	5.01 ± 7.15	30.8 ± 4.82	34.97 ± 0.65	33.5 ± 6.3
GDS2547	648.55 ± 255.09	1.82 ± 4.24	23.7 ± 5.86	27.61 ± 0.75	27.3 ± 6.63
GDS4431	43.35 ± 284.56	0 ± 0	0 ± 0	14.31 ± 0.41	2.55 ± 3.82
GDS5218	4207.29 ± 1589.79	7.96 ± 4.08	24.0 ± 5.09	30.70 ± 0.38	31.5 ± 6.15
GDS3966	2976.15 ± 811.92	10.6 ± 7.83	31.7 ± 4.71	36.77 ± 0.54	37.9 ± 6.44
GDS4185	2.92 ± 40.20	0 ± 0	0 ± 0	8.95 ± 0.79	0.701 ± 3.11
GDS2737	0.79 ± 7.33	0 ± 0	0 ± 0	6.13 ± 0.63	1.73 ± 2.82
GDS4719	0.36 ± 1.88	0 ± 0	0 ± 0	7.99 ± 1.18	2.8 ± 5.31

Table 3.3: Results for the SNP datasets when using all of the instances. Reproducibility Scores are shown in the form of mean ± standard deviation.

Name	#instances (Majority %)	#features	#biomarkers	uRS %	mRS %	oRS %
GSE15826	164 (54%)	909549	0	0 ± 0	2.03 ± 0.06	2.5 ± 0.58
GSE25103 [41]	122 (92%)	908512	325	0.27 ± 0.14	1.53 ± 0.08	3.94 ± 2.22
GSE25104 [32], [41]	122 (92%)	909547	326	0.27 ± 0.14	1.45 ± 0.03	3.94 ± 2.22
GSE18333 [34]	82 (54%)	909606	0	0 ± 0	0 ± 0	0 ± 0
GSE15096 [37]	69 (58%)	909457	106482	5.37 ± 2.9	24.83 ± 4.35	33.4 ± 5.28
GSE15097 [37]	68 (59%)	909456	108224	4.9 ± 2.9	23.59 ± 2.61	34.2 ± 5.64
GSE13429 [55]	39 (79%)	262314	1267	1.66 ± 0.44	9.18 ± 0.51	21.4 ± 3.76

Table 3.4: Results for the SNP datasets when using half of the instances. Reproducibility Scores and average number of biomarkers are shown in the form of mean ± standard deviation.

Name	Average #biomarkers	uRS %	$\widehat{RS}$ %	pRS %	oRS %
GSE15826	1.21 ± 5.41	0 ± 0	2.03 ± 0.06	1.2 ± 0.47	1.21 ± 0.79
GSE25103	309.61 ± 78.24	0.13 ± 0.11	0.27 ± 0.14	1.36 ± 0.12	6.63 ± 6.79
GSE25104	309.93 ± 78.14	0.13 ± 0.11	0.27 ± 0.14	1.60 ± 0.07	6.64 ± 6.8
GSE18333	0.01 ± 0.1	0 ± 0	0 ± 0	0.18 ± 0.23	0 ± 0
GSE15096	22393.87 ± 20624.09	0.74 ± 0.60	5.37 ± 2.9	6.89 ± 3.74	16.1 ± 6.91
GSE15097	22788.62 ± 23342.37	0.77 ± 0.62	4.9 ± 2.99	12.16 ± 6.1	16.8 ± 6.85
GSE13429	339.29 ± 170.11	1.73 ± 0.44	1.66 ± 0.44	5.12 ± 0.78	14 ± 5.04

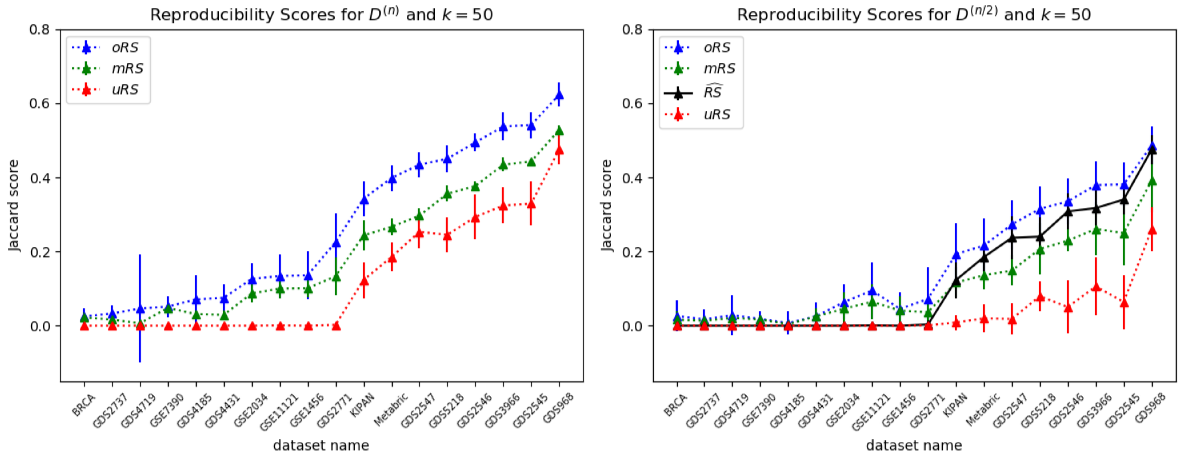

Figure 3.3: Reproducibility scores (mean and standard deviation) for all 16 continuous datasets, both for complete datasets with n instances (left) and for half-sized with $\frac{n}{2}$ instances (right), for $k = 50$ iterations. The x-axes (for both plots) are sorted by the value of the over-bound for the $D^{(n)}$ datasets. We see, in both, that the over-bound oRS is consistently higher than the under-bound uRS, and the middle-value estimate mRS is between them. Moreover, the right plot shows that the "truth" $\widehat{RS}$ is also between uRS and oRS.

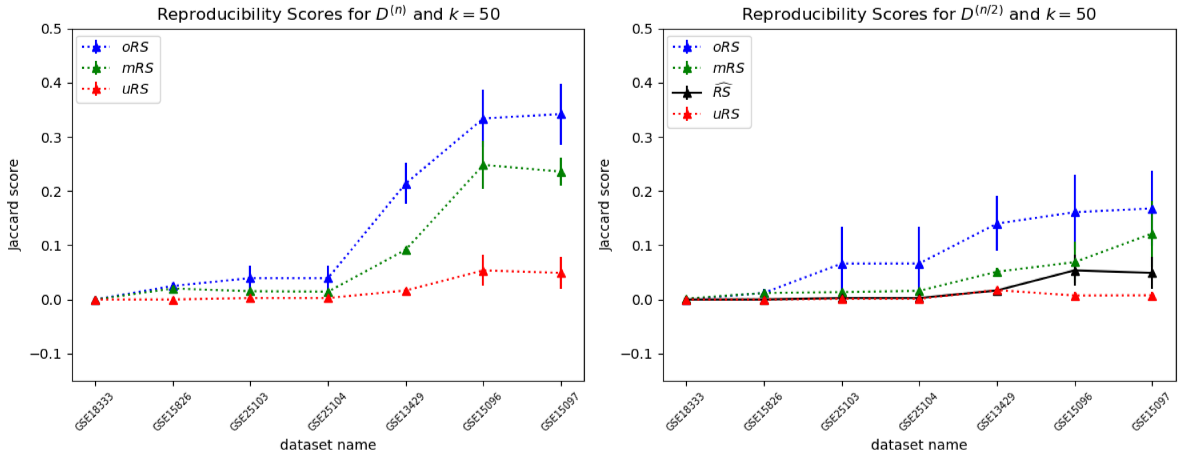

Figure 3.4: Reproducibility scores (mean and standard deviation) for 7 SNP datasets, both for complete datasets with n instances (left) and for half-sized with $\frac{n}{2}$ instances (right), for 50 iterations. The x-axes (for both plots) is sorted by the value of the overbound oRS for the $D^{(n)}$ datasets. We see, in both, that the over-bound oRS is consistently higher than the under-bound uRS, and the middle-value estimate mRS is between them. Moreover, the right plot shows that the "truth" $\widehat{RS}$ is also within the range of oRS and uRS.

Chapter 4

Discussion

4.1 Future Work

While the message of this dissertation is very general, the specific analyses all used the standard discovery algorithm $\text{BD}_{t,0.05,BH}$. Our empirical studies all dealt with standard datasets, whose values were either all real values or all categorical values; none had some of each. We only considered datasets whose labels are binary and our use of t-test implicitly assumes they are Gaussian; see also footnote 6. Our analytic model considers the overlap of biomarkers found from two datasets, of the same size. (That is, we do not consider how the biomarkers obtained from a 100-element dataset, overlap with those from a 300-element dataset.) Finally, our analysis estimated the expected RS, for a given BD and dataset D. It would be interesting to explore a variant of this: Given a dataset D and a minimum score $s > 0$, find the "$\text{BD}'(D,\ s)$ discovery algorithm" that would produce the biomarker set whose expected Jaccard score would be at least s. (This might mean adjusting the p-value cut-off, and/or including some specific FDR algorithm, or some other modification.)

4.2 Contributions

There are effective ways to accurately estimate the reproducibility of the biomarker set obtained from a (labeled) dataset and Bioaarker Discovery method. This dissertation provides (1) a formal definition of the reproducibility of the biomarker set obtained from a (labeled) dataset and Biomarker Discovery method, (2) techniques that accurately estimate this reproducibility score, and (3) empirical results that demonstrate the effectiveness of those techniques, for a range of t-test based BD methods over 25 real datasets.

No Machine Learning paper simply presents a learned classifier – they always accompany that classifier with a claim about its accuracy (eg, "87 +- 2% accuracy", or "AUC of 0.72" or ...). Similarly, we anticipate that no future Biomarker Discovery paper will simply present a set of biomarkers. Instead, future biomarker discovery researchers will always accompany that set with

some falsifiable claim – if not based on a biological experiment, or its use in a downstream classifier, then about its reproducibility. This dissertation has (1) motivated and defined this reproducibility framework – in terms of the *set* of biomarkers, produced from a labeled dataset and a specific biomarker discovery method, evaluated in terms of its expected Jaccard score wrt similarly generated datasets, (2) presented a body of specific algorithms for effectively bounding this reproducibility score for BD's that are based on t-test, and (3) demonstrated that this algorithms do effectively bound that RS.

References

[1] A. Abraham, M. P. Milham, A. Di Martino, R. C. Craddock, D. Samaras, B. Thirion, and G. Varoquaux, "Deriving reproducible biomarkers from multi-site resting-state data: An autism-based example," *NeuroImage*, vol. 147, pp. 736–745, 2017. 2

[2] A. A. Alizadeh, A. J. Gentles, A. J. Alencar, C. L. Liu, H. E. Kohrt, R. Houot, M. J. Goldstein, S. Zhao, Y. Natkunam, R. H. Advani, *et al.*, "Prediction of survival in diffuse large b-cell lymphoma based on the expression of two genes reflecting tumor and microenvironment," *Blood*, blood–2011, 2011. 2

[3] M. D. Alter, R. Kharkar, K. E. Ramsey, D. W. Craig, R. D. Melmed, T. A. Grebe, R. C. Bay, S. Ober-Reynolds, J. Kirwan, J. J. Jones, *et al.*, "Autism and increased paternal age related changes in global levels of gene expression regulation," *PloS one*, vol. 6, no. 2, e16715, 2011. 20

[4] A. M. Becker, K. H. Dao, B. K. Han, R. Kornu, S. Lakhanpal, A. B. Mobley, Q.-Z. Li, Y. Lian, T. Wu, A. M. Reimold, *et al.*, "Sle peripheral blood b cell, t cell and myeloid cell transcriptomes display unique profiles and each subset contributes to the interferon signature," *PloS one*, vol. 8, no. 6, e67003, 2013. 20

[5] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: A practical and powerful approach to multiple testing," *Journal of the royal statistical society. Series B (Methodological)*, pp. 289–300, 1995. 12, 34

[6] Y. Benjamini and D. Yekutieli, "The control of the false discovery rate in multiple testing under dependency," *Annals of statistics*, pp. 1165–1188, 2001. 34

[7] N. Blau, "Genetics of phenylketonuria: Then and now," *Human mutation*, vol. 37, no. 6, pp. 508–515, 2016. 1

[8] A.-L. Boulesteix and M. Slawski, "Stability and aggregation of ranked gene lists," *Briefings in bioinformatics*, vol. 10, no. 5, pp. 556–568, 2009. 1

[9] R. O. Burney, S. Talbi, A. E. Hamilton, K. C. Vo, M. Nyegaard, C. R. Nezhat, B. A. Lessey, and L. C. Giudice, "Gene expression analysis of endometrium reveals progesterone resistance and candidate susceptibility genes in women with endometriosis," *Endocrinology*, vol. 148, no. 8, pp. 3814–3826, 2007. 20

[10] W. S. Bush and J. H. Moore, "Genome-wide association studies," *PLoS computational biology*, vol. 8, no. 12, e1002822, 2012. 1

[11] U. R. Chandran, C. Ma, R. Dhir, M. Bisceglia, M. Lyons-Weiler, W. Liang, G. Michalopoulos, M. Becich, and F. A. Monzon, "Gene expression profiles of prostate cancer reveal involvement of multiple molecular pathways in the metastatic process," *BMC cancer*, vol. 7, no. 1, p. 64, 2007. 20

[12] C. Curtis, S. P. Shah, S.-F. Chin, G. Turashvili, O. M. Rueda, M. J. Dunning, D. Speed, A. G. Lynch, S. Samarajiwa, Y. Yuan, *et al.*, "The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups," *Nature*, vol. 486, no. 7403, p. 346, 2012. 20

[13] *Data generated by the tcga research network: Http://cancergenome.nih.gov/*, Downloaded Feb 2019. 20

[14] C. Desmedt, F. Piette, S. Loi, Y. Wang, F. Lallemand, B. Haibe-Kains, G. Viale, M. Delorenzi, Y. Zhang, M. S. d'Assignies, *et al.*, "Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the transbig multicenter independent validation series," *Clinical cancer research*, vol. 13, no. 11, pp. 3207–3214, 2007. 3, 20

[15] R. Dhami, M. A. Passini, and E. H. Schuchman, "Identification of novel biomarkers for niemann–pick disease using gene expression analysis of acid sphingomyelinase knockout mice," *Molecular Therapy*, vol. 13, no. 3, pp. 556–564, 2006. 2

[16] O. J. Dunn, "Multiple comparisons among means," *Journal of the American statistical association*, vol. 56, no. 293, pp. 52–64, 1961. 34

[17] B. Efron, "Bootstrap methods: Another look at the jackknife," in *Breakthroughs in statistics*, Springer, 1992, pp. 569–593. 11

[18] L. Ein-Dor, I. Kela, G. Getz, D. Givol, and E. Domany, "Outcome signature genes in breast cancer: Is there a unique set?" *Bioinformatics*, vol. 21, no. 2, pp. 171–178, 2004. 2, 6, 8

[19] L. Ein-Dor, O. Zuk, and E. Domany, "Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer," *Proceedings of the National Academy of Sciences*, vol. 103, no. 15, pp. 5923–5928, 2006. 6, 8

[20] D. R. Fernandez, T. Telarico, E. Bonilla, Q. Li, S. Banerjee, F. A. Middleton, P. E. Phillips, M. K. Crow, S. Oess, W. Muller-Esterl, *et al.*, "Activation of mammalian target of rapamycin controls the loss of tcrζ in lupus t cells through hres-1/rab4-regulated lysosomal degradation," *The Journal of Immunology*, vol. 182, no. 4, pp. 2063–2073, 2009. 20

[21] M. Gormley, W. Dampier, A. Ertel, B. Karacali, and A. Tozeren, "Prediction potential of candidate biomarker sets identified and validated on gene expression data from multiple datasets," *BMC bioinformatics*, vol. 8, no. 1, p. 415, 2007. 2

[22] L. Guo, E. K. Lobenhofer, C. Wang, R. Shippy, S. C. Harris, L. Zhang, N. Mei, T. Chen, D. Herman, F. M. Goodsaid, *et al.*, "Rat toxicogenomic study reveals analytical consistency across microarray platforms," *Nature biotechnology*, vol. 24, no. 9, p. 1162, 2006. 8, 9

[23] A. M. Gustafson, R. Soldi, C. Anderlind, M. B. Scholand, J. Qian, X. Zhang, K. Cooper, D. Walker, A. McWilliams, G. Liu, *et al.*, "Airway pi3k pathway activation is an early and reversible event in lung cancer development," *Science translational medicine*, vol. 2, no. 26, 26ra25–26ra25, 2010. 20

[24] M. Haubitz, D. M. Good, A. Woywodt, H. Haller, H. Rupprecht, D. Theodorescu, M. Dakna, J. J. Coon, and H. Mischak, "Identification and validation of urinary biomarkers for differential diagnosis and evaluation of therapeutic intervention in anti-neutrophil cytoplasmic antibody-associated vasculitis," *Molecular & Cellular Proteomics*, vol. 8, no. 10, pp. 2296–2307, 2009. 2

[25] Y. Hochberg, "A sharper bonferroni procedure for multiple tests of significance," *Biometrika*, vol. 75, no. 4, pp. 800–802, 1988. 34

[26] S. Holm, "A simple sequentially rejective multiple test procedure," *Scandinavian journal of statistics*, pp. 65–70, 1979. 34

[27] H. Hotelling, "Analysis of a complex of statistical variables into principal components.," *Journal of educational psychology*, vol. 24, no. 6, p. 417, 1933. 4

[28] J. Hutcheson, J. C. Scatizzi, A. M. Siddiqui, G. K. Haines III, T. Wu, Q.-Z. Li, L. S. Davis, C. Mohan, and H. Perlman, "Combined deficiency of proapoptotic regulators bim and fas results in the early onset of systemic autoimmunity," *Immunity*, vol. 28, no. 2, pp. 206–217, 2008. 20

[29] J. P. Ioannidis, "Biomarker failures," *Clinical chemistry*, vol. 59, no. 1, pp. 202–204, 2013. 6

[30] L. Klebanov, X. Qiu, S. Welle, and A. Yakovlev, "Statistical methods and microarray data," *Nature biotechnology*, vol. 25, no. 1, p. 25, 2007. 9

[31] R. J. Koenig, C. M. Peterson, R. L. Jones, C. Saudek, M. Lehrman, and A. Cerami, "Correlation of glucose regulation and hemoglobin aic in diabetes mellitus," *New England Journal of Medicine*, vol. 295, no. 8, pp. 417–420, 1976. 1

[32] C.-H. Lee, J. S.-M. Chang, S.-H. Syu, T.-S. Wong, J. Y.-W. Chan, Y.-C. Tang, Z.-P. Yang, W.-C. Yang, C.-T. Chen, S.-C. Lu, *et al.*, "Il-1β promotes malignant transformation and tumor aggressiveness in oral cancer," *Journal of cellular physiology*, vol. 230, no. 4, pp. 875–884, 2015. 21

[33] M. Li, G. Hong, J. Cheng, J. Li, H. Cai, X. Li, Q. Guan, M. Tong, H. Li, and Z. Guo, "Identifying reproducible molecular biomarkers for gastric cancer metastasis with the aid of recurrence information," *Scientific reports*, vol. 6, p. 24 869, 2016. 6, 8

[34] X. Mao, Y. Yu, L. K. Boyd, G. Ren, D. Lin, T. Chaplin, S. C. Kudahetti, E. Stankiewicz, L. Xue, L. Beltran, *et al.*, "Distinct genomic alterations in prostate cancers in chinese and western populations suggest alternative pathways of prostate carcinogenesis," *Cancer research*, vol. 70, no. 13, pp. 5207–5212, 2010. 21

[35] G. L. G. Miklos and R. Maleszka, "Microarray reality checks in the context of a complex disease," *Nature biotechnology*, vol. 22, no. 5, p. 615, 2004. 2

[36] H. Mischak, G. Allmaier, R. Apweiler, T. Attwood, M. Baumann, A. Benigni, S. E. Bennett, R. Bischoff, E. Bongcam-Rudloff, G. Capasso, *et al.*, "Recommendations for biomarker identification and qualification in clinical proteomics," *Science translational medicine*, vol. 2, no. 46, 46ps42–46ps42, 2010. 2, 6

[37] E. Närvä, R. Autio, N. Rahkonen, L. Kong, N. Harrison, D. Kitsberg, L. Borghese, J. Itskovitz-Eldor, O. Rasool, P. Dvorak, *et al.*, "High-resolution dna analysis of human embryonic stem cell lines reveals culture-induced copy number changes and loss of heterozygosity," *Nature biotechnology*, vol. 28, no. 4, p. 371, 2010. 21

[38] P. Patil, P.-O. Bachant-Winner, B. Haibe-Kains, and J. T. Leek, "Test set bias affects reproducibility of gene signatures," *Bioinformatics*, vol. 31, no. 14, pp. 2318–2323, 2015. 20

[39] Y. Pawitan, J. Bjöhle, L. Amler, A.-L. Borg, S. Egyhazi, P. Hall, X. Han, L. Holmberg, F. Huang, S. Klaar, *et al.*, "Gene expression profiling spares early breast cancer patients from adjuvant therapy: Derived and validated in two population-based cohorts," *Breast cancer research*, vol. 7, no. 6, R953, 2005. 20

[40] J. Pearl, *Causality*. Cambridge university press, 2009. 2

[41] C.-H. Peng, C.-T. Liao, S.-C. Peng, Y.-J. Chen, A.-J. Cheng, J.-L. Juang, C.-Y. Tsai, T.-C. Chen, Y.-J. Chuang, C.-Y. Tang, *et al.*, "A novel molecular signature identified by systems genetics approach predicts prognosis in oral squamous cell carcinoma," *PloS one*, vol. 6, no. 8, e23452, 2011. 21

[42] S. Ramaswamy, K. N. Ross, E. S. Lander, and T. R. Golub, "A molecular signature of metastasis in primary solid tumors," *Nature genetics*, vol. 33, no. 1, p. 49, 2002. 2

[43] U. Raue, T. A. Trappe, S. T. Estrem, H.-R. Qian, L. M. Helvering, R. C. Smith, and S. W. Trappe, "Transcriptome signature of resistance exercise adaptations: Mixed muscle and fiber type specific profiles in young and old adults," *American Journal of Physiology-Heart and Circulatory Physiology*, 2012. 20

[44] K. E. Rieger, W.-J. Hong, V. G. Tusher, J. Tang, R. Tibshirani, and G. Chu, "Toxicity from radiation therapy associated with abnormal transcriptional responses to dna damage," *Proceedings of the National Academy of Sciences*, vol. 101, no. 17, pp. 6635–6640, 2004. 20

[45] K. Rossing, H. Mischak, M. Dakna, P. Zürbig, J. Novak, B. A. Julian, D. M. Good, J. J. Coon, L. Tarnow, P. Rossing, *et al.*, "Urinary proteomics in diabetes and ckd," *Journal of the American Society of Nephrology*, vol. 19, no. 7, pp. 1283–1290, 2008. 2

[46] J. M. Satagopan, E. Venkatraman, and C. B. Begg, "Two-stage designs for gene–disease association studies with sample size constraints," *Biometrics*, vol. 60, no. 3, pp. 589–597, 2004. 4

[47] M. Schmidt, D. Böhm, C. Von Törne, E. Steiner, A. Puhl, H. Pilch, H.-A. Lehr, J. G. Hengstler, H. Kölbl, and M. Gehrmann, "The humoral immune system has a key prognostic impact in node-negative breast cancer," *Cancer research*, vol. 68, no. 13, pp. 5405–5413, 2008. 20

[48] L. Shi, L. H. Reid, W. D. Jones, R. Shippy, J. A. Warrington, S. C. Baker, P. J. Collins, F. De Longueville, E. S. Kawasaki, K. Y. Lee, *et al.*, "The microarray quality control (maqc) project shows inter-and intraplatform reproducibility of gene expression measurements," *Nature biotechnology*, vol. 24, no. 9, p. 1151, 2006. 8, 9

[49] T. Shlomi, M. N. Cabili, and E. Ruppin, "Predicting metabolic biomarkers of human inborn errors of metabolism," *Molecular systems biology*, vol. 5, no. 1, p. 263, 2009. 2

[50] T. Sørlie, C. M. Perou, R. Tibshirani, T. Aas, S. Geisler, H. Johnsen, T. Hastie, M. B. Eisen, M. Van De Rijn, S. S. Jeffrey, *et al.*, "Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications," *Proceedings of the National Academy of Sciences*, vol. 98, no. 19, pp. 10 869–10 874, 2001. 2

[51] A. Spira, J. E. Beane, V. Shah, K. Steiling, G. Liu, F. Schembri, S. Gilman, Y.-M. Dumas, P. Calner, P. Sebastiani, *et al.*, "Airway epithelial gene expression in the diagnostic evaluation of smokers with suspect lung cancer," *Nature medicine*, vol. 13, no. 3, p. 361, 2007. 20

[52] K. Strimbu and J. A. Tavel, "What are biomarkers?" *Current Opinion in HIV and AIDS*, vol. 5, no. 6, p. 463, 2010. 1

[53] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, *et al.*, "Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles," *Proceedings of the National Academy of Sciences*, vol. 102, no. 43, pp. 15 545–15 550, 2005. 4

[54] 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, *et al.*, "Gene expression profiling predicts clinical outcome of breast cancer," *nature*, vol. 415, no. 6871, p. 530, 2002. 1, 2

[55] R. Venkatachalam, E. T. Verwiel, E. J. Kamping, E. Hoenselaar, H. Görgens, H. K. Schackert, J. H. J. van Krieken, M. J. Ligtenberg, N. Hoogerbrugge, A. G. van Kessel, *et al.*, "Identification of candidate predisposing copy number variants in familial and early-onset colorectal cancer patients," *International journal of cancer*, vol. 129, no. 7, pp. 1635–1642, 2011. 21

[56] Y. Wang, J. G. Klijn, Y. Zhang, A. M. Sieuwerts, M. P. Look, F. Yang, D. Talantov, M. Timmermans, M. E. Meijer-van Gelder, J. Yu, *et al.*, "Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer," *The Lancet*, vol. 365, no. 9460, pp. 671–679, 2005. 1, 2, 20

[57] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016. 5, 10

[58] L. Xu, S. S. Shen, Y. Hoshida, A. Subramanian, K. Ross, J.-P. Brunet, S. N. Wagner, S. Ramaswamy, J. P. Mesirov, and R. O. Hynes, "Gene expression changes in an animal melanoma model correlate with aggressiveness of human melanoma metastases," *Molecular Cancer Research*, vol. 6, no. 5, pp. 760–769, 2008. 20

[59] Y. P. Yu, D. Landsittel, L. Jing, J. Nelson, B. Ren, L. Liu, C. McDonald, R. Thomas, R. Dhir, S. Finkelstein, *et al.*, "Gene expression alterations in prostate cancer predicting tumor aggression and preceding development of malignancy," *Journal of clinical oncology*, vol. 22, no. 14, pp. 2790–2799, 2004. 20

[60] H. U. Zacharias, T. Rehberg, S. Mehrl, D. Richtmann, T. Wettig, P. J. Oefner, R. Spang, W. Gronwald, and M. Altenbuchinger, "Scale-invariant biomarker discovery in urine and plasma metabolite fingerprints," *Journal of proteome research*, vol. 16, no. 10, pp. 3596–3605, 2017. 2

[61] M. Zhang, C. Yao, Z. Guo, J. Zou, L. Zhang, H. Xiao, D. Wang, D. Yang, X. Gong, J. Zhu, *et al.*, "Apparently low reproducibility of true differential expression discoveries in microarray studies," *Bioinformatics*, vol. 24, no. 18, pp. 2057–2063, 2008. 6, 8

[62] J. Zou, C. Hao, G. Hong, J. Zheng, L. He, and Z. Guo, "Revealing weak differential gene expressions and their reproducible functions associated with breast cancer metastasis," *Computational biology and chemistry*, vol. 39, pp. 1–5, 2012. 2, 8, 18–20, 36

Appendix A

Exploring other settings

For consistency, all of the experiments in the main text used the same $\text{BD}_{t,0.05,BH}$ biomarker discovery algorithm. It is worth noting that this tool produces very different numbers of biomarkers over the 25 datasets we considered; see Figure A.1, which also relates this to the number of features in each dataset.

However, there are many other approaches that can, and have, been used in other association studies. Here, we continue to consider only the t-test as the main statistical significance test. Appendix A.1 explores different options for the p-value threshold and the p-value adjustment method, to see how changing these affect the reproducibility results. This dissertation introduced three different approximations for the *Reproducibility Score* – uRS, mRS, oRS. Appendix A.2 explores how these approximations change as we adjust the number of iterations of running the algorithms, k.

A.1 p-value adjustment methods and p-value threshold

Note that FDR-correction is designed to increase precision (aka positive predictive value), but it might reduce recall (aka true positive rate or sensitivity), which means it might reduce the Jaccard score, and so lead to lower RS scores. We therefore experimented with different correction methods to see their effect on the reproducibility of biomarker sets.

Figure A.2 shows the effect of Benjamini+Hochberg (BH) FDR correction on the reproducibility scores, across all datasets. We see that this FDR correc-

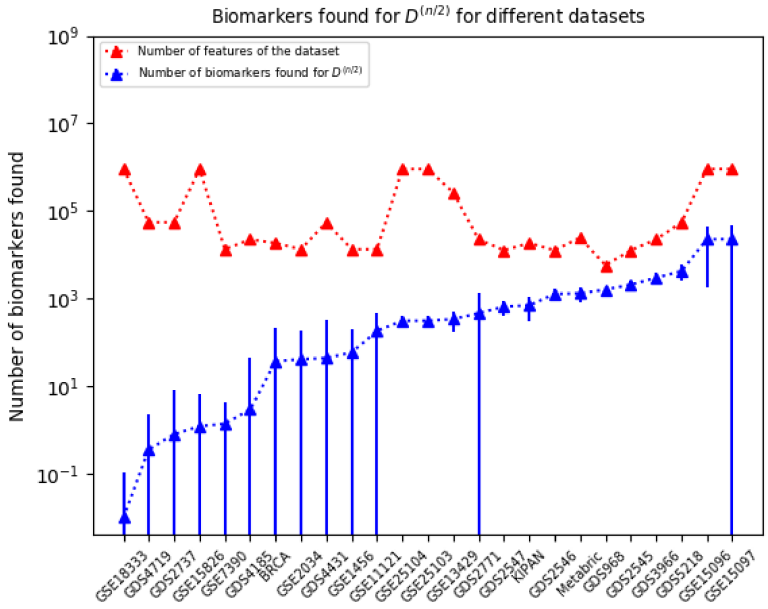

Figure A.1: Number of biomarkers (mean ± sd) found for $D^{(n/2)}$ when using $\text{BD}_{t,0.05,BH}$ over $k = 20$ iterations for various datasets, compared to the number of features in each dataset. Note the y-axis is a log-scale.

tion is *detrimental*, as it reduces the reproducibility scores across all datasets: While it is designed to reduce false discoveries (and hence increase precision), this may mean it is reducing recall, which collectively leads to a smaller Jaccard score.

There are many other methods for reducing FDR, in addition to Benjamini+Hochberg (BH) [5], including: Benjamini+Yekutieli (BY) [6], Bonferroni [16], Hochberg [25] and Holm [26]. Figure A.3 shows the results of these 5 FDR methods, as well as the "no-FDR" approach, over all 25 datasets – here showing $\widehat{RS}$ wrt the half-datasets $D^{(n/2)}$; see Equation 2.7. We see again that "no-FDR" remains the best approach, and that BH is the 2nd best, followed by the others.

We also anticipate the RS score will depend on the p-value threshold used to determine the significance of each feature – *i.e.*, $\text{BD}_{t,\tau,BH}$, for various $\tau \in (0, 0.1)$. While most studies use a threshold of $\tau = 0.05$, this number is fairly arbitrary. Here we explored how the RS changed with different values of τ. Figure A.4 shows that the reproducibility score for $D^{(n/2)}$ appears monotonic

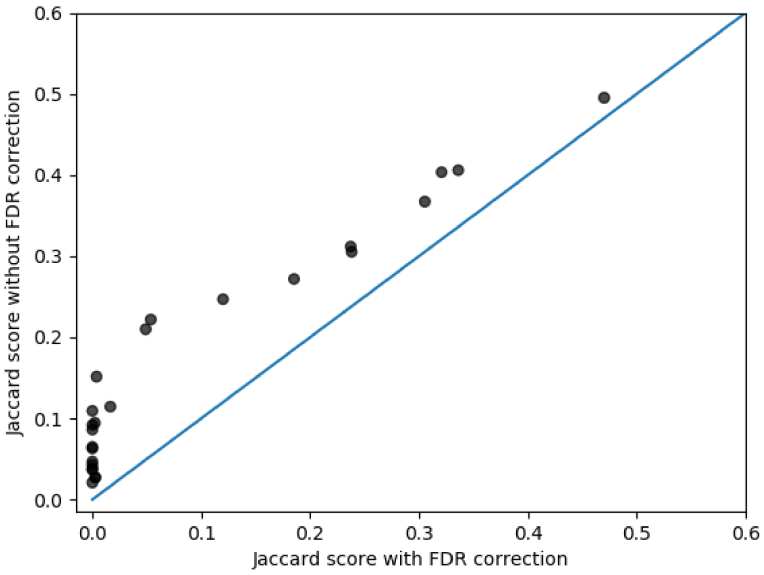

Figure A.2: Scatter plot of Reproducibility Scores for all 25 datasets: each (x, y) point represents the average $D^{(n/2)}$ Jaccard scores for a single dataset (using disjoint subset pairs), where the x-value represents the score with FDR correction ($\text{BD}_{t,0.05,BH}$) and the y-value which represents the score without FDR correction ($\text{BD}_{t,0.05,-}$). A point above the diagonal line means the FDR correction led to inferior performance.

with τ – larger τ produces higher RS.

A.2 Changing k

Our various approximation algorithms each use k, the number of samples used by the algorithms. As we often work with large datasets, these algorithms can be very time consuming (even though they have been optimized), motivating us to explore how these algorithms scale, based on this parameter.

We ran these algorithms for our largest dataset, Metabric, but varied the number of iterations. Figure A.5 shows that we obtained very similar results, whether we used $k = 10$, up to $k = 50$. We also looked at the MRS values for all 25 datasets when running the algorithm for $k = 10$ iterations versus $k = 50$ iterations and the difference between the two is very close to 0 for most cases; see Figure A.6. We ran a paired t-test for these values, which resulted in a

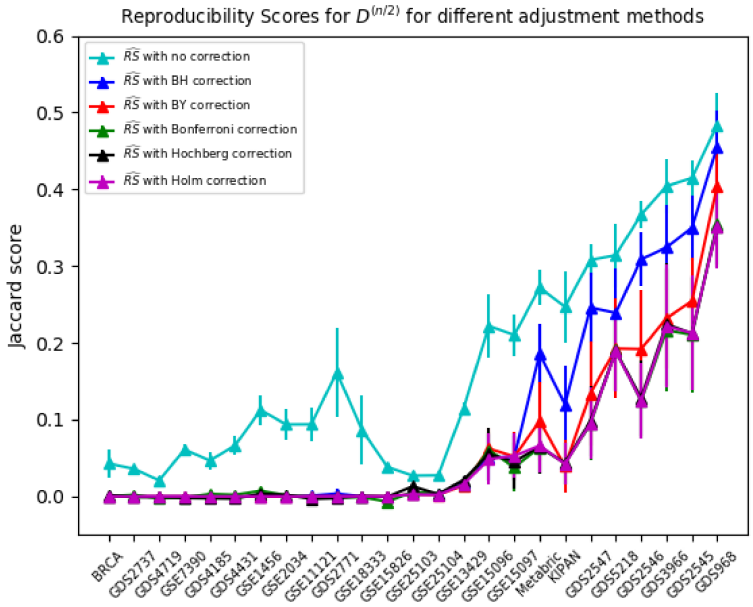

Figure A.3: Reproducibility scores for datasets when using $D^{(n/2)}$ for different p-value adjustment methods – *i.e.*, $\text{BD}_{t,0.05,\chi}$ for 5 different FDR adjustment methods χ, including BH and "no".

p-value of 0.91 that tells us that MRS with $k = 50$ and MRS with $k = 50$ are not statistically different.

A.3 PO Score

We have used Jaccard score as our similarity measure throughout all of our experiments. This is a symmetric measure, meaning it provides information about a pair of datasets $\{ A, B \}$ where $J(A, B) = J(B, A)$. It can be very useful when comparing the results from different experiments or evaluating the outcome when trying to replicate results from a previous study. However, there are other options for the similarity measure that are not symmetric and can be provided together with the set of biomarkers for each dataset. One of these options is the PO score, which is used by the Zou *et al.* [62] meta-study: for each (ordered) pair of biomarker sets $[B_i, B_j]$,

$$PO[B_i, B_j] \quad = \quad \frac{|B_i \cap B_j|}{|B_i|} \times 100\%. \tag{A.1}$$

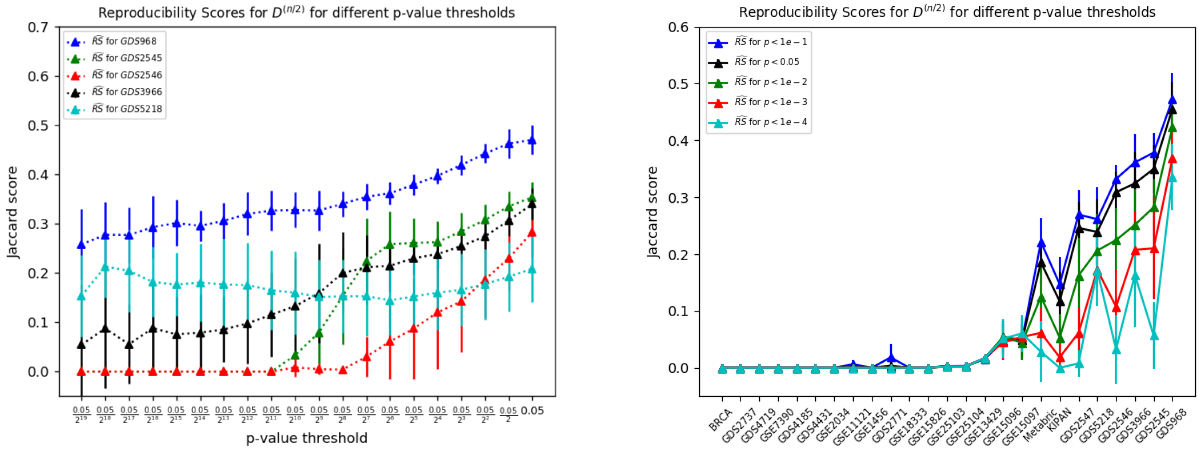

Figure A.4: Reproducibility scores for various datasets when using $D^{(n/2)}$ for different p-value thresholds – $\text{BD}_{t,\tau,BH}$, for various $\tau \in (0, 0.1)$.

(Notice this is an assymmetric variant of the Jaccard score [Equation 2.2].) (As that paper also reported the number of biomarkers found for each dataset, we could recover the associated Jaccard score.)

Unlike the Jaccard score, this measure can be reported alongside a set of biomarkers and claim that a certain number of these biomarkers should be replicated, if another study is to be done with the same criteria for the data.

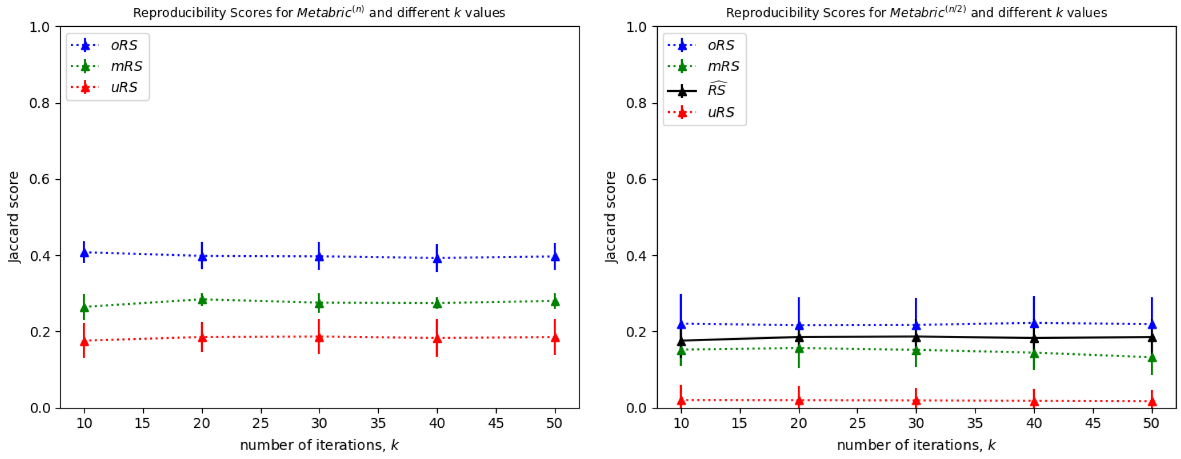

Figure A.5: Reproducibility scores for different numbers of iterations, for the Metabric dataset when using half the data.

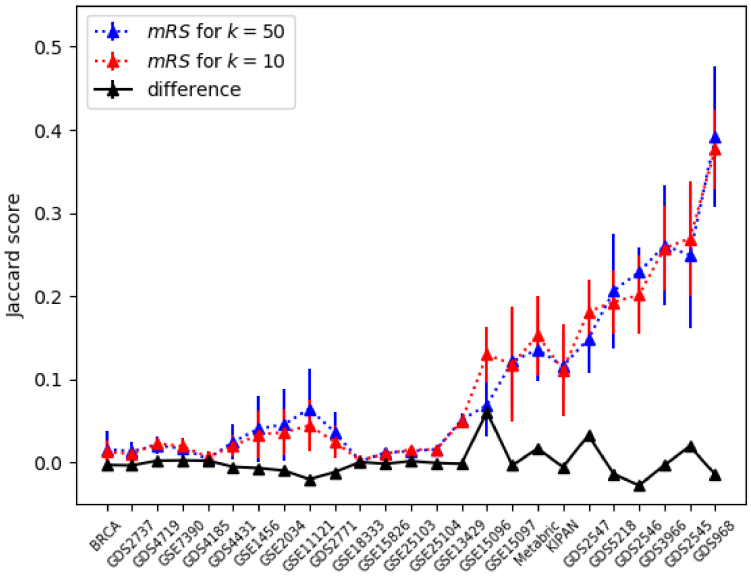

Figure A.6: MRS values for all 25 datasets when running $k = 10$ versus $k = 50$ iterations on $D^{(n/2)}$ subsets.